DOI:
10.1145/3301315

Document Version
Peer reviewed version

Metastability of the Logit Dynamics for Asymptotically Well-Behaved Potential Games*

Diodato Ferraioli† Carmine Ventre‡

Abstract

Convergence rate and stability of a solution concept are classically measured in terms of "eventually" and "forever", respectively. In the wake of recent computational criticisms to this approach, we study whether these time frames can be updated to have states computed "quickly" and stable for "long enough".

Logit dynamics allows irrationality in players' behavior, and may take time exponential in the number of players n to converge to a stable state (i.e., a certain distribution over pure strategy profiles). We prove that every potential game, for which the behavior of the logit dynamics is not chaotic as n increases, admits distributions stable for a super-polynomial number of steps in n no matter the players' irrationality, and the starting profile of the dynamics. The convergence rate to these *metastable distributions* is polynomial in n when the players are not too rational.

Our proofs build upon the new concept of *partitioned Markov chains*, that might be of independent interest, and a number of involved technical contributions.

Keywords: Logit Dynamics; Bounded Rationality; Markov chains; Approximate Equilibria; Potential Games

*An extended abstract of this work appeared in the proceedings of MFCS 2015 [19]. The work has also been accepted as a contributed talk at GAMES 2016.

†Dipartimento di Ingegneria dell'Informazione ed Elettrica e Matematica Applicata Università di Salerno, Italy. E-mail: dferraioli@unisa.it.

‡CSEE, University of Essex, UK. E-mail: C.Ventre@essex.ac.uk.

1 Introduction

One of the most prominent assumptions in game theory dictates that people are rational. This is contrasted by many concrete instances of people making irrational choices in certain strategic situations, such as stock markets [35]. This might be due to the incapacity of exactly determining one's own utilities: the strategic game is played with utilities perturbed by some noise.

Logit dynamics [7] incorporates this noise in players' actions and then is advocated to be a good model for people behavior. In more detail, logit dynamics features a rationality level $\beta \geq 0$ (equivalently, a noise level $1/\beta$) and each player is assumed to play a strategy with a probability which is proportional to her corresponding utility and β. So the higher β is, the less noise there is and the more rational players are. Logit dynamics can then be seen as a noisy best-response dynamics.

The natural equilibrium concept for logit dynamics is defined by a probability distribution over the pure strategy profiles of the game. Whilst for best-response dynamics pure Nash equilibria are stable states, in logit dynamics there is a chance, which is inversely proportional to β, that players deviate from such strategy profiles. Pure Nash equilibria are then not an adequate solution concept for this dynamics. However, the random process defined by the logit dynamics can be modeled via an ergodic Markov chain. Stability in Markov chains is represented by the concept of stationary distributions. These distributions, dubbed logit equilibria, are suggested as a suitable solution concept in this context due to their properties [4]. For example, from the results known in Markov chain literature, we know that every game possesses a logit equilibrium and that this equilibrium is unique. The absence of either of these guarantees is often considered a weakness of pure Nash equilibria. Nevertheless, as for Nash equilibria, the computation of logit equilibria may be computationally hard depending on whether the chain mixes rapidly (i.e., in time polynomial in the number of players) or not [3].

As the hardness of computing Nash equilibria justifies approximate notions of the concept [26, 10], so Auletta et al. [5] look at an approximation of logit equilibria that they call *metastable distributions*. These distributions aim to *describe* regularities arising during the transient phase of Markov chains before stationarity has been reached. Indeed, they are distributions that remain stable for a time which is long enough for the observer rather than forever. Roughly speaking, the stability of the distributions in this concept is measured in terms of the generations living some historical era, while stationary distributions remain stable throughout all the generations. When the convergence to logit equilibria is too slow, then there are generations which are outlived by the computation of the stationary distribution. For these generations, metastable distributions[1] grant an otherwise impossible descriptive power. (We refer the interested reader to [5] for a complete overview of the rationale of metastability and for examples of this concept.) Nevertheless, it is unclear whether and which strategic games enable the logit dynamics to possess these distributions and if they are quickly reached.

The focus of this paper is the study of metastable distributions for the logit dynamics run on the class of potential games [28]. Potential games are an important and widely studied class of games modeling many strategic settings. Each such game satisfies a number of appealing properties, the existence of pure Nash equilibria being one of them. A general study of metastability of potential games was left open by [5] and assumes particular interest due to the known hardness results, see e.g. [17], which suggest that the computation of pure Nash equilibria for them is an intractable problem, even for centralized algorithms.

Our contribution. Our main result proves that every n-player potential game has a metastable distribution for *each* starting profile of the logit dynamics, when the behavior of the logit dynamics on the game is not chaotic in n. These distributions remain stable for a time which is super-polynomial in n, if one is content with being within distance $\varepsilon > 0$ from the distributions. (The distance is defined in

[1] Observe that the concept of metastable distribution is defined for every Markov chain, not just for the logit dynamics. However, the latter has been its main application.

this context as the total variation distance, see below.) We also prove that the convergence rate to these distributions, called *pseudo-mixing time*, is polynomial in n for values of β not too big when compared to the (inverse of the) maximum difference in potential of neighboring profiles. Note that when β is very high then logit dynamics is "close" to the best-response dynamics and therefore it is impossible to prove in general quick convergence results for potential games due to the aforementioned hardness results. We then give a picture which is as complete as possible. (To maintain n as our only parameter of interest, we assume that the logarithm of the number of strategies available to players is upper bounded by a polynomial in n; this assumption can, however, be relaxed to prove bounds asymptotic in n and in the logarithm of the maximum number of strategies.)

We remark that our results are *asymptotic* in n, which imposes some requirement on the potential games of interest to avoid a chaotic behavior (in n) of the logit dynamics run on a game; a wild behavior of the Markov chain would not allow any meaningful asymptotic guarantee for a game.

To understand the kind of requirement that we need, it will be useful to liken our objective to the following: we are given a graph algorithm and we are asked to find for which graphs the algorithm runs in time polynomial in v, the number of nodes. Clearly, an answer to this question must consider not a single graph, for which no asymptotic result makes sense, but a sequence of graphs indexed by the number of nodes, as, for example, cliques, rings, stars, complete binary trees, and so on. Moreover, one also requires that these sequences are well defined for infinitely many values of v. Indeed, if arbitrary sequences of graphs were allowed, then it would be easy to adversarially build sequences that for infinitely many values of v use graphs for which the algorithm runs in polynomial time, whereas for other infinitely many values of v they use graphs for which the algorithm requires exponential time. For this sequence of graphs it is not possible to bound the running time with a polynomial, and thus we reach the wrong conclusion that no graph exists for which the algorithm is polynomial. This explains why in most of the literature about graph algorithms one focuses on a "well-defined" sequence of graphs (i.e., sequences of graphs for which the value of v does not alter the structure of the graph).

Here, the object of our analysis is the Markov chain modeling the evolution of the logit dynamics for a given potential game. As for algorithms on graphs, in order to prove asymptotic results, we need to take in account asymptotic well-defined sequences of these objects. Our definition considers (a sequence of) potential games for which we can describe, for every n sufficiently large, the time needed to leave a subset of profiles as either polynomial or super-polynomial. We call these games *asymptotically well-behaved*. Generally speaking, this is the class of games for which the kind of asymptotic results we are after are meaningful. Indeed, if we were unable to say if the dynamics leaves a subset of states in polynomial or in super-polynomial time (e.g., if for infinitely many values of n we can describe this time with a polynomial, and for infinitely many other values we need a super-polynomial to this aim) then, depending on the (metastable) distribution of interest, we may be unable to prove asymptotic results on the behavior of the dynamics starting from this subset of states. Our main result can be restated to say that super-polynomial stability and polynomial convergence from every starting profile of the logit dynamics holds for *every potential game for which it makes sense to prove these asymptotic results*.[2]

Moreover, we stress that similar assumptions are made in related literature on logit dynamics either implicitly (as in [29, 3], where it is assumed that certain properties of the potential function do not change as n changes), or explicitly, by considering specific games that clearly enjoy this property [5]. Moreover, asymptotic results on the mixing time of Markov chains do require some assumption on the behavior of the chain (technically, the minimum bottleneck ratio must either be a polynomial or a super-polynomial) usually implicitly guaranteed by the definition of the chain at hand. Given that our objective is much more complex than bounding the mixing time (i.e., measuring asymptotically the transient phase of the chain – defined on *a* potential game – and ascertain stability of *and* convergence time to metastable distributions) a similar, yet stronger, requirement ought to be used.

[2] In Appendix B we give a route to our definition, showing how less stringent definitions lead to ad-hoc n-player "chaotic" potential games that do not admit metastable distributions with super-polynomially long stability and polynomial convergence.

Our Technique. The proof of the above results consists of two main steps. We first devise, in Section 3 a sufficient property for every Markov chain to have, for every starting state, a distribution that is metastable for a large number of steps and reached quickly. The main idea behind this sufficient condition is that when the dynamics starts from a subset from which it is "hard to leave" and in which it is "easy to mix", then the dynamics will stay for a long time close to the stationary distribution restricted to that subset. Moreover, if a subset is "easy-to-leave," then the dynamics will quickly reach a "hard-to-leave" subset. The sufficient property, named *partitioned Markov chain*, intuitively consists a partition of the states into subsets that are asymptotically "hard-to-leave & easy-to-mix" or "easy-to-leave".

The second step amounts to showing that for every asymptotically well-behaved potential game the logit dynamics admits such a partition. The proof of this result, given in Section 4, builds on a number of technical contributions, some of which might be of independent interest. They mainly concern Markov chains. The concepts of interest are mixing time (how long the chain takes to mix), bottleneck ratio (intuitively, how hard it is for the stationary distribution to leave a subset of states), hitting time (how long the chain takes to hit a certain subset of states) and spectral properties of the transition matrix of Markov chains. In particular, we define a procedure which computes the required partition for these games. We iteratively identify in the set of pure strategy profiles the "hard-to-leave" subsets. To prove that these subsets are "easy-to-mix", we firstly relate the pseudo-mixing time to the mixing time of a certain family of restricted Markov chains. We then prove that the mixing time of these chains is polynomial by using a spectral characterization of the transition matrix of restricted Markov chains. Finally, the proof that the remaining profiles are "easy-to-leave" mainly relies on a connection between bottleneck ratio and hitting time. Specifically, we prove both an upper bound and a lower bound on the hitting time of a subset of states in terms of the bottleneck ratio of its complement.

The tool of partitioned Markov chains also allows us to simplify the proof of asymptotic metastability for specific classes of games of interests. We exemplify this in Section 4.3 where we apply our results to three specific games: the Curie-Weiss game, a Pigou-like congestion game and the opinion formation game on complete bipartite graphs [18]. Incidentally, we essentially close an open problem of [5] about metastability of the Curie-Weiss game. We remark that these games encompass all the classes of games for which the behavior of logit dynamics has been studied in the literature.

While we introduce the tool of partitioned Markov chains to prove our main result, we believe that it may be of independent interest in the analysis of the behavior of more general Markov chains. To highlight this aspect of our work, we discuss in Section 5, a couple of problems, namely graph clustering algorithms and exponential random network generation, that may benefit from the perspective and tools of asymptotic metastability and partitioned Markov chains.

In appendix, we further complement the above contributions. We indeed prove additional spectral results about the transition matrix of Markov chains defined by logit dynamics for a strategic (not necessarily potential) game (cf. Section D). These results enhance our understanding of the dynamics and pave the way to further advancements in the area.

Related works. Blume [7] introduced logit dynamics for modeling a noisy-rational behavior in game dynamics. Early works about this dynamics have focused on its long-term behavior: Blume [7] showed that, for 2×2 coordination games and potential games, the long-term behavior of the system is concentrated around a specific Nash equilibrium; Alòs-Ferrer and Netzer [1] gave a general characterization of long-term behavior of logit dynamics for wider classes of games. Several works gave bounds on the time that the dynamics takes to reach specific Nash equilibria of a game: Ellison [16] considered graphical coordination games on cliques and rings; Peyton Young [33] and Montanari and Saberi [29] extended this work to more general families of graphs. Different class of games have been considered by Asadpour and Saberi [2] that focused on a class of congestion games, and by Ferraioli et al. [18] that instead considered discrete preference games. Auletta et al. [4] were the first to propose the stationary distribution of the logit dynamics Markov chain as a new equilibrium concept in game theory and to

focus on the time the dynamics takes to get close to this equilibrium [3]. The logit response function has also been used for defining another equilibrium concept, known as *quantal response equilibrium* [27]. This differs from the logit equilibrium since it is a product distribution (like Nash equilibrium).

In physics, chemistry, and biology, metastability is a phenomenon related to the evolution of systems under noisy dynamics. In particular, metastability concerns moves between regions of the state spaces and the existence of multiple, well separated time scales: at short time scales, the system appears to be in a quasi-equilibrium, but really explores only a confined region of the available space state, while, at larger time scales, it undergoes transitions between such different regions. Previous research about metastability aims at expressing typical features of a metastable state and to evaluate the transition time between metastable states. Several monographs on the subject are available in literature (see, for example, [20, 31, 8, 21]). Auletta et al. [5] applied metastability to probability distributions, introducing the concepts of metastable distribution and pseudo-mixing time and proving results for some specific potential games.

Roughly speaking, metastability is a kind of approximation for stationarity. From this point of view, metastable distributions may be likened to approximate equilibria. Two different approaches to approximated equilibria have been proposed in literature. In the multiplicative version [10] a profile is an approximate equilibrium as long as each player gains at least a factor $(1-\varepsilon)$ of the payoff she gets by playing any other strategy: these equilibria have been shown to be computationally hard both in general [13] and for congestion games [36]. In the additive version [23], a profile is an approximate equilibrium as long as each player gains at least the payoff she gains by playing any other strategy minus a small additive factor $\varepsilon > 0$: for these equilibria a quasi-polynomial time approximation scheme exists [26] but it is impossible to have an FPTAS [9].

2 Preliminary definitions

A (discrete-time) *Markov chain* $\mathcal{M} = (S, P)$ is a sequence of random variables $\{X_i \colon i \in \mathbb{N}\}$ such that $X_i \in S$ for every $i \in \mathbb{N}$ and $\mathbf{P}\left(X_i = \mathbf{y} \mid X_{i-1} = \mathbf{x}\right) = \mathbf{P}_{\mathbf{x}}\left(X_1 = \mathbf{y}\right)$[3] $= P(\mathbf{x}, \mathbf{y})$ for every $\mathbf{x}, \mathbf{y} \in S$ and every $i \in \mathbb{N}$. Here, S is the (finite) *state space* of the Markov chain, and $P \colon S \times S \to [0, 1]$ is its *transition matrix*. We will denote with P^t the t-step transition matrix that describes the probability $\mathbf{P}\left(X_i = \mathbf{y} \mid X_{i-t} = \mathbf{x}\right)$ of the state of the chain after t steps given the current state.

A Markov chain is *irreducible* if for every pair of states $\mathbf{x}, \mathbf{y} \in S$, there is t such that $P^t(\mathbf{x}, \mathbf{y}) > 0$, i.e., there is a chance to eventually reach $\mathbf{y}$ from every starting state. A Markov chain is *aperiodic* if for every pair of states $\mathbf{x}, \mathbf{y} \in S$, the greatest common denominator of the t's such that $P^t(\mathbf{x}, \mathbf{y}) > 0$ is 1, i.e., the chain can eventually visit $\mathbf{y}$ at any time and not only in a specific subset of times. A Markov chain with finite state space is *ergodic* if it is irreducible and aperiodic. In particular, this implies that there is $t^\star$ such that $P^{t^\star}(\mathbf{x}, \mathbf{y}) > 0$ for every $\mathbf{x}, \mathbf{y} \in S$. If a Markov chain is ergodic, then from every initial state x the distribution $P^t(\mathbf{x}, \cdot)$ over states of S will eventually converge to a *stationary distribution* π, such that $\pi P = \pi$.

A Markov chain with transition matrix P and state space S is said to be *reversible* with respect to a distribution π if, for all $\mathbf{x}, \mathbf{y} \in S$, it holds that $\pi(\mathbf{x})P(\mathbf{x}, \mathbf{y}) = \pi(\mathbf{y})P(\mathbf{y}, \mathbf{x})$. If an ergodic chain is reversible with respect to π, then π is its stationary distribution. Therefore when this happens, to simplify our exposition we simply say that the matrix P is reversible.

Given a set of states L we let $\overline{L}$ and ∂L to denote, respectively, its complementary set, i.e., $\overline{L} = S \setminus L$, and the border of L, that is the set of profiles in L with at least a neighbor in $\overline{L}$. Moreover, we say that a set L is *connected* if, for every $\mathbf{x}, \mathbf{y} \in L$, there are $\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_k \in L$ with $\mathbf{x}_0 = \mathbf{x}$, $\mathbf{x}_k = \mathbf{y}$ and $P(\mathbf{x}_{i-1}, \mathbf{x}_i) > 0$ for each $i = 1, \ldots, k$.

[3] Throughout this work, we denote with $\mathbf{P}_{\mathbf{x}}(\cdot)$ the probability of an event conditioned on the starting state of the Markov chain being $\mathbf{x}$.

2.1 Convergence of Markov chains

Mixing time. Arguably, the principal notion to measure the rate of convergence of a Markov chain to its stationary distribution is the *mixing time*, which is defined as follows. Let us set

$$d(t) = \max_{\mathbf{x} \in S} \left\| P^t(\mathbf{x}, \cdot) - \pi \right\|_{\mathrm{TV}},$$

where the *total variation distance* $\|\mu - \nu\|_{\mathrm{TV}}$ between two probability distributions μ and ν on the same state space S is defined as

$$\|\mu - \nu\|_{\mathrm{TV}} = \max_{A \subset S} |\mu(A) - \nu(A)| = \frac{1}{2} \sum_{\mathbf{x} \in S} |\mu(\mathbf{x}) - \nu(\mathbf{x})|,$$

where $\mu(A) = \sum_{\mathbf{x} \in A} \mu(\mathbf{x})$ and $\nu(A) = \sum_{\mathbf{x} \in A} \nu(\mathbf{x})$. For $0 < \varepsilon < 1/2$, the mixing time of the logit dynamics is defined as

$$t_{\mathrm{mix}}(\varepsilon) = \min\{t \in \mathbb{N} \colon d(t) \leq \varepsilon\}.$$

It is usual to set $\varepsilon = 1/4$ or $\varepsilon = 1/2e$. We write t_{mix} to mean $t_{\mathrm{mix}}(1/4)$ and we refer generically to "mixing time" when the actual value of ε is immaterial. Observe that $t_{\mathrm{mix}}(\varepsilon) \leq \left\lceil \log_2 \varepsilon^{-1} \right\rceil t_{\mathrm{mix}}$.

Relaxation time. Another important measure of convergence for Markov chains is given by the *relaxation time*. Let P be the transition matrix of a Markov chain with finite state space S; let us label the eigenvalues of P in non-increasing order

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{|S|}.$$

It is well-known (see, for example, Lemma 12.1 in [25]) that $\lambda_1 = 1$ and, if P is ergodic, then $\lambda_2 < 1$ and $\lambda_{|S|} > -1$. We set $\lambda^\star$ as the largest eigenvalue in absolute value other than λ_1, i.e.,

$$\lambda^\star = \max_{i=2,\ldots,|S|} \{|\lambda_i|\}.$$

The relaxation time t_{rel} of a Markov chain $\mathcal{M}$ is defined as

$$t_{\mathrm{rel}} = \frac{1}{1 - \lambda^\star}.$$

The relaxation time is related to the mixing time by the following theorem (see, for example, Theorems 12.3 and 12.4 in [25]).

Theorem 2.1. *Let P be the transition matrix of an ergodic and reversible Markov chain with state space S and stationary distribution π. Then*

$$(t_{\mathrm{rel}} - 1) \log 2 \leq t_{\mathrm{mix}} \leq \log \left(\frac{4}{\pi_{\mathrm{min}}} \right) t_{\mathrm{rel}},$$

where $\pi_{\mathrm{min}} = \min_{\mathbf{x} \in S} \pi(\mathbf{x})$.

Hitting time. In some cases, we are interested in bounding the first time that the chain hits a profile in a certain set of states, also known as its *hitting time*. Formally, for a set $L \subseteq S$, we denote by τ_L the random variable denoting the hitting time of L. Note that the hitting time, differently from mixing and relaxation time, depends on where the dynamics starts.

Bottleneck ratio. Quite central in our study is the concept of *bottleneck ratio*. Consider an ergodic Markov chain with finite state space S, transition matrix P, and stationary distribution π. The probability distribution $Q(\mathbf{x}, \mathbf{y}) = \pi(\mathbf{x})P(\mathbf{x}, \mathbf{y})$ is of particular interest and is sometimes called the *edge stationary distribution*. Note that if the chain is reversible then $Q(\mathbf{x}, \mathbf{y}) = Q(\mathbf{y}, \mathbf{x})$. For every $L \subseteq S$, $L \neq \emptyset$, we let $Q(L, S \setminus L) = \sum_{\mathbf{x}\in L, \mathbf{y}\in S\setminus L} Q(\mathbf{x}, \mathbf{y})$. Then the bottleneck ratio of L is

$$B(L) = \frac{Q(L, S \setminus L)}{\pi(L)}.$$

We use the following theorem to derive lower bounds to the mixing time (see, for example, Theorem 7.3 in [25]).

Theorem 2.2. *Let $\mathcal{M} = \{X_t \colon t \in \mathbb{N}\}$ be an ergodic Markov chain with state space S, transition matrix P, and stationary distribution π. Let $L \subseteq S$ be any set with $\pi(L) \leq 1/2$. Then the mixing time is*

$$t_{\text{mix}} \geq \frac{1}{4B(L)}.$$

The bottleneck ratio is also strictly related to the relaxation time. Indeed, let

$$B_\star = \min_{L\colon \pi(R)\leq 1/2} B(L).$$

Then the following theorem holds (see, for example, Theorem 13.14 in [25]).

Theorem 2.3. *Let P be the transition matrix of an ergodic and reversible Markov chain with state space S. Let λ_2 be the second largest eigenvalue of P. Then*

$$\frac{B_\star^2}{2} \leq 1 - \lambda_2 \leq 2B_\star.$$

2.2 Metastability

In this section we give formal definitions of *metastable distributions* and *pseudo-mixing time*. We also survey some of the tools used for our results. For a more detailed description we refer the reader to [5].

Definition 2.1. Let P be the transition matrix of a Markov chain with state space S. A probability distribution μ over S is $(\varepsilon, \mathcal{T})$-*metastable*, with $\varepsilon > 0$ and $\mathcal{T} \in \mathbb{N}$, for P if for every $0 \leq t \leq \mathcal{T}$ it holds that

$$\left\|\mu P^t - \mu\right\|_{\text{TV}} \leq \varepsilon.$$

The definition of metastable distribution captures the idea of a distribution that behaves approximately like the stationary distribution: if we start from such a distribution and run the chain we stay close to it for a "long" time. Some interesting properties of metastable distributions are discussed in [5], including the following lemmata, that turn out to be useful for proving our results.

Lemma 2.1 ([5])**.** *Let P be a Markov chain with finite state space S and stationary distribution π. For a subset of states $L \subseteq S$ let π_L be the stationary distribution conditioned on L, i.e.*

$$\pi_L(\mathbf{x}) = \begin{cases} \pi(\mathbf{x})/\pi(L), & \text{if } \mathbf{x} \in L; \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

Then, π_L is $(B(L), 1)$-metastable.

Lemma 2.2 ([5])**.** *If μ is $(\varepsilon, 1)$-metastable for P then μ is $(\varepsilon\mathcal{T}, \mathcal{T})$-metastable for P.*

Among all metastable distributions, we are interested in the ones that are quickly reached from a (possibly large) set of states. This motivates the following definition.

Definition 2.2. Let P be the transition matrix of a Markov chain with state space S, let $L \subseteq S$ be a non-empty set of states and let μ be a probability distribution over S. We define the *pseudo-mixing time* $t_\mu^L(\varepsilon)$ as

$$t_\mu^L(\varepsilon) = \inf\{t \in \mathbb{N}\colon \left\|P^t(\mathbf{x}, \cdot) - \mu\right\|_{\text{TV}} \leq \varepsilon \text{ for all } \mathbf{x} \in L\}.$$

Since the stationary distribution π of an ergodic Markov chain is reached within ε in time $t_{\text{mix}}(\varepsilon)$ from every state, according to Definition 2.2 we have that $t_\pi^S(\varepsilon) = t_{\text{mix}}(\varepsilon)$. The following simple lemma connects metastability and pseudo-mixing time.

Lemma 2.3 ([5])**.** *Let μ be a $(\varepsilon_1, \mathcal{T})$-metastable distribution and let $L \subseteq S$ be a set of states such that $t_\mu^L(\varepsilon_2)$ is finite. Then for every $\mathbf{x} \in L$ it holds that $\left\|P^t(\mathbf{x}, \cdot) - \mu\right\|_{\text{TV}} \leq \varepsilon_1 + \varepsilon_2$ for every $t_\mu^L(\varepsilon_2) \leq t \leq t_\mu^L(\varepsilon_2) + \mathcal{T}$.*

2.3 Asymptotic Metastability

The notions and results about metastability and pseudo-mixing time introduced above apply to a single Markov chain. Auletta et al. [5] adopted these notions to evaluate the asymptotic behavior of a parametrized class of Markov chains, where the parameter n is the logarithm of the number of states. Specifically, they consider a *sequence* $\mathbf{M}$ *of n-chains*, $\mathcal{M}_n$ being the unique Markov chain in the sequence with a number of states whose logarithm is n, and they analyze the asymptotic properties of these Markov chains. Since we do not have a single Markov chain but a sequence of them, one for each n, we need to consider an asymptotic counterpart of the notions above. Auletta et al. [5], in fact, showed that the special Markov chains defined in their work enjoys the following property, that we name *asymptotic metastability*.

Definition 2.3. Let $\mathbf{M}$ be a sequence of n-chains. Let C be a class of functions in n closed under multiplication[4]. We say that $\mathbf{M}$ is *asymptotically metastable* for C if for every $\varepsilon > 0$ there are functions $p = p_\varepsilon \in C$ and $q = q_\varepsilon \notin C$ such that for each n sufficiently large, the Markov chain $\mathcal{M}_n$ converges in time at most $p(n)$ from each starting state to a $(\varepsilon, q(n))$-metastable distribution.

In other words, for each starting state the Markov chain "quickly" converges to a distribution that remains metastable for an amount of time that is "much larger" than the time one takes to reach it, where "quickly" and "much larger" are asymptotically measured on the number of states.

3 Partitioned Markov Chains Are Asymptotically Metastable

We next give a sufficient property for *every* ergodic Markov chain to be asymptotically metastable. Specifically, in this section we will introduce the concept of *partitioned* Markov chain. Then, we prove that partitioned Markov chain are asymptotically metastable.

Note that we will next focus only on ergodic Markov chains whose mixing time is not in the class C, otherwise the stationary distribution enjoys the desired properties of stability and convergence.

[4]Clearly, different definitions of C give rise to more or less interesting distributions, depending on the application of interest. We will mainly use C to divide polynomials from superpolynomials, except for some of the applications outside game theory in Section 5 where C will be defined to separate logarithmic functions from superlogarithmic ones.

3.1 Partitioned Markov Chains

Let $\mathbf{M}$ be a sequence of n-chains. Let P be the transition matrix of $\mathcal{M}_n$, for some $n > 0$, and let π be the corresponding stationary distribution. For non-empty $L \subseteq S$, we define a Markov chain with state space L and transition matrix $\mathring{P}_L$ defined as follows.

$$\mathring{P}_L(\mathbf{x},\mathbf{y}) = \begin{cases} P(\mathbf{x},\mathbf{y}) & \text{if } \mathbf{x} \neq \mathbf{y}; \\ 1 - \sum_{\substack{\mathbf{z}\in L,\\ \mathbf{z}\neq\mathbf{x}}} P(\mathbf{x},\mathbf{z}) = P(\mathbf{x},\mathbf{x}) + \sum_{\mathbf{z}\in S\setminus L} P(\mathbf{x},\mathbf{z}) & \text{otherwise.} \end{cases} \tag{2}$$

It easy to check that the stationary distribution of this Markov chain is given by the distribution $\pi_L(\mathbf{x}) = \frac{\pi(\mathbf{x})}{\pi(L)}$, for every $\mathbf{x} \in L$. Note also that the Markov chain defined upon $\mathring{P}_L$ is aperiodic, since the Markov chain defined upon P is, and it will be irreducible if L is a connected set. For a fixed $\varepsilon > 0$, we will denote with $t^L_{\text{mix}}(\varepsilon)$ the mixing time of the chain described in (2). We also denote with $B_L(A)$ the bottleneck ratio of $A \subset L$ in the Markov chain with state space L and transition matrix $\mathring{P}_L$.

We are now ready to introduce the definition of partitioned Markov chain.

Definition 3.1. Let $\mathbf{M}$ be a sequence of n-chains. We say that $\mathbf{M}$ is *partitioned* for a class C of functions closed under multiplication if for every $\varepsilon > 0$ there is $p = p_\varepsilon \in C$ and $q = q_\varepsilon \notin C$ such that for each n there is a family of connected subsets $R_1, \ldots, R_k$ of the state space S of $\mathcal{M}_n$, with $k \geq 1$, and a partition $T_1, \ldots, T_k, N$ of S, with $T_i \subseteq R_i$ for every $i = 1, \ldots, k$, such that

1. the bottleneck ratio of R_i is at most $1/q(n)$, for every $i = 1, \ldots, k$;
2. the mixing time $t^{R_i}_{\text{mix}}(\varepsilon)$ is at most $p(n)$, for every $i = 1, \ldots, k$;
3. for every $i = 1, \ldots, k$ and for every $\mathbf{x} \in T_i$, it holds that
$$\mathbf{P}_\mathbf{x}\left(\tau_{S\setminus R_i} \leq t^{R_i}_{\text{mix}}(\varepsilon)\right) \leq \varepsilon;$$
4. for every $\mathbf{x} \in N$, it holds that
$$\mathbf{P}_\mathbf{x}\left(\tau_{\bigcup_i T_i} \leq p(n)\right) \geq 1 - \varepsilon.$$

Note that we allow in the above definition that T_i, for some $i = 1, \ldots, k$, or N are empty. Linking back to the intuition discussed in the introduction, $R_1, \ldots, R_k$ represent the "easy-to-mix" subsets of states (condition 2); these sets play a crucial role in defining distributions that are metastable for very long time (condition 1). However, when the Markov chain starts close to the boundary of some R_i, it is likely to leave R_i quickly. Since we are interested in "easy-to-mix & hard-to-leave" subsets of states, for each R_i we identify its *core* T_i as the set of states from which the Markov chain takes long time to leave R_i (condition 3). The distinction between core and non-core states will help in proving that metastable distributions are quickly reached from every starting state.

The main result of this section proves that a partitioned Markov chain is asymptotically metastable.

Theorem 3.1. *Let $\mathbf{M}$ be a sequence of n-chains. If $\mathbf{M}$ is partitioned for C, then $\mathbf{M}$ is asymptotically metastable for C.*

3.2 Proof of Theorem 3.1

A high level idea of the proof is discussed next. We initially need to define the metastable distributions to which the Markov chain converges. To describe the distributions of interest, we leverage known results connecting bottleneck ratio and metastability. In particular, it turns out that the stationary distribution of

the chain restricted to of a subset of states with bottleneck ratio at most the inverse of a super-polynomial, as defined in (1), is metastable for a super-polynomial amount of time (Lemma 2.1). In this way we can easily "build" metastable distributions from the sets R_i given by the definition of partitioned chain (cf. Proposition 3.1).

What about the pseudo-mixing time of this distribution? We distinguish two cases. First we consider states that are in the "*core*" of the support of this distribution, namely the sets T_i given by the definition of partitioned chain. We show that the pseudo-mixing time from these states is related to the mixing time of the chain restricted to R_i as described in (2) (see Corollary 3.1). Then, being the mixing time of the chains restricted to R_i polynomial by construction, it follows that the pseudo-mixing time from the core is polynomial.

What about out-of-core states? Suppose that there is a state from which the chain takes long time to converge to a metastable distribution. Then it must be the case that the chain takes long time to hit the core of one such distribution with high probability. However, this cannot be the case since, by definition, the partitioned chain from every non-core state quickly hits a state in the core of some distribution.

Next we formally prove Theorem 3.1. In particular, the proof follows from the Propositions 3.1, 3.2 and 3.3 given below that, respectively, describe the metastable distributions, bound the pseudo-mixing time from the core, and evaluate the behavior of the Markov chain when starting from non-core states.

Identifying the metastable distributions. We start by proving that some distributions defined on the sets R_i are metastable for long time.

Proposition 3.1. *Let $\mathbf{M}$ be a sequence of n-chains. If $\mathbf{M}$ is partitioned for C, then for every $\varepsilon > 0$ there exists a function $\mathcal{T} = \mathcal{T}_\varepsilon \notin C$ such that for each[5] i and each n, the distribution μ_i that sets $\mu_i(\mathbf{x}) = \pi(\mathbf{x})/\pi(R_i)$, where π is the stationary distribution of $\mathcal{M}_n$, is $(\varepsilon, \mathcal{T}(n))$-metastable.*

Proof. Fix i. Given $\varepsilon > 0$, consider the function $\mathcal{T} = \mathcal{T}_\varepsilon$ such that $\mathcal{T}(n) = \frac{\varepsilon}{B(R_i)} \geq \varepsilon q(n)$, where R_i is the support of μ_i. By the definition of q, $\mathcal{T} \notin C$.

By Lemma 2.1, μ_i is $(B(R_i), 1)$-metastable. By Lemma 2.2, μ_i is also $(B(R_i) \cdot \mathcal{T}(n), \mathcal{T}(n))$-metastable. The lemma follows since $B(R_i) \cdot \mathcal{T}(n) = \varepsilon$. □

Finally, the following lemma shows that a combination of metastable distributions is metastable.

Lemma 3.1. *Let P the transition matrix of a Markov chain with state space S and let μ_i be a distribution $(\varepsilon_i, \mathcal{T}_i)$-metastable for P, for $i = 1, 2, \ldots$. Set $\varepsilon = \max_i \varepsilon_i$ and $\mathcal{T} = \min_i\{\mathcal{T}_i\}$. Then, the distribution $\mu = \sum_i \alpha_i \mu_i$, with $\sum_i \alpha_i = 1$ and $\alpha_i \geq 0$, is $(\varepsilon, \mathcal{T})$-metastable.*

Proof. For every $t \leq \mathcal{T}$ we have

$$\begin{aligned}
\left\|\mu P^t - \mu\right\|_{\mathrm{TV}} &= \max_{A \subseteq S} \left|(\mu P^t)(A) - \mu(A)\right| \\
&= \max_{A \subseteq S} \left| \sum_i \alpha_i \left((\mu_i P^t)(A) - \mu_i(A)\right) \right| \\
&\leq \sum_i \alpha_i \max_{A \subseteq S} \left|(\mu_i P^t)(A) - \mu_i(A)\right| \leq \varepsilon.
\end{aligned}$$

□

[5]There may be values of n for which the partition uses less than i "components" and thus R_i, T_i and μ_i are not well defined. However, asymptotic bounds on the metastability and the pseudo-mixing time of μ_i are well defined as long as there are infinite values of n for which R_i is given. Moreover, since every partition contains at least one "component" for every input, we have that there exists n_0 such that for every $i \leq \max_{n \geq n_0} k(n)$, R_i is defined infinite times.

Pseudo-mixing time from the core. Now we prove that the a partitioned Markov chain converges quickly to the metastable distribution μ_i defined above, whenever the starting point is selected from the core T_i of this distribution. Specifically, we prove the following proposition.

Proposition 3.2. *Let $\mathbf{M}$ be a sequence of n-chains and fix $\varepsilon > 0$. If $\mathbf{M}$ is partitioned for C, then there is a function $p_\star \in C$ such that for each i and each n, the pseudo-mixing time of μ_i from T_i is $t_{\mu_i}^{T_i}(2\varepsilon) = O(p_\star(n))$.*

In order to prove Proposition 3.2, consider the Markov chain defined in (2). Let us abuse the notation and denote with $\mathring{P}_L$ and π_L also the Markov chain and the distribution defined on the entire state space S, with $\mathring{P}_L(\mathbf{x}, \mathbf{y}) = 0$ if $\mathbf{x} \notin L$ or $\mathbf{y} \notin L$, and similarly $\pi_L(\mathbf{x}) = 0$ when $\mathbf{x} \notin L$.

Recall that ∂L is the border of L, that is the set of profiles in L with at least a neighbor in $S \setminus L$. Recall that $\tau_{S\setminus L}$ is the random variable denoting the first time the Markov chain with transition matrix P hits a state $\mathbf{x} \in S \setminus L$. The following lemma formally proves the intuitive fact that, by starting from a state in L the chain P and the chain $\mathring{P}_L$ are the same up to the time in which the former chain hits a state in $S \setminus L$. The proof uses the well-known coupling technique (cf., e.g., [25]) which is summarized in Appendix A.1.

Lemma 3.2. *Let P be the transition matrix of a Markov chain with state space S and let $\mathring{P}_L$ be the restriction of P to $L \subseteq S$, $L \neq \emptyset$, as given in* (2)*. Then, for every $\mathbf{x} \in L$ and for every $t > 0$,*

$$\left\| P^t(\mathbf{x}, \cdot) - \mathring{P}_L^t(\mathbf{x}, \cdot) \right\|_{\mathrm{TV}} \leq \mathbf{P}_{\mathbf{x}} \left(\tau_{S\setminus L} \leq t \right).$$

Proof. Consider the following coupling $(X_t, Y_t)_{t>0}$ of the Markov chains with transition matrix P and $\mathring{P}_L$, respectively:

- If $X_i = Y_i \in L \setminus \partial L$, then we update the first chain according to P and obtain X_{i+1}; we then set $Y_{i+1} = X_{i+1}$;
- If $X_i = Y_i \in \partial L$, then we update the first chain according to P: if $X_{i+1} \in L$, then we set $Y_{i+1} = X_{i+1}$, otherwise we set $Y_{i+1} = Y_i$;
- If $X_i \neq Y_i$, then we update the chains independently.

Since $X_0 = Y_0 = \mathbf{x} \in L$, we have that $X_t \neq Y_t$ only if $\tau_{S\setminus L} \leq t$. Thus, by the properties of couplings (see Theorem A.1), we have

$$\left\| P^t(\mathbf{x}, \cdot) - \mathring{P}_L^t(\mathbf{x}, \cdot) \right\|_{\mathrm{TV}} \leq \mathbf{P}_{\mathbf{x}} \left(X_t \neq Y_t \right) \leq \mathbf{P}_{\mathbf{x}} \left(\tau_{S\setminus L} \leq t \right). \qquad \square$$

The following corollary follows from the Lemma 3.2 and the triangle inequality property of the total variation distance.

Corollary 3.1. *Let P the transition matrix of a Markov chain with state space S and let $\mathring{P}_L$ be the restriction of P to a non-empty $L \subseteq S$ as given in* (2)*. Then, for every $\mathbf{x} \in L$ and for every $t > 0$,*

$$\left\| P^t(\mathbf{x}, \cdot) - \pi_L \right\|_{\mathrm{TV}} \leq \left\| \mathring{P}_L^t(\mathbf{x}, \cdot) - \pi_L \right\|_{\mathrm{TV}} + \mathbf{P}_{\mathbf{x}} \left(\tau_{S\setminus L} \leq t \right).$$

Using Corollary 3.1 we can prove Proposition 3.2.

Proof of Proposition 3.2. Fix n. For each $\mathbf{x} \in T_i$, by Corollary 3.1 and since $\mathbf{P}_{\mathbf{x}} \left(\tau_{S\setminus R_i} \leq t_{\mathrm{mix}}^{R_i}(\varepsilon) \right) \leq \varepsilon$, we obtain

$$\left\| P^{t_{\mathrm{mix}}^{R_i}(\varepsilon)}(\mathbf{x}, \cdot) - \mu_i \right\|_{\mathrm{TV}} \leq \varepsilon + \varepsilon.$$

The lemma follows from the observation that $t_{\mathrm{mix}}^{R_i}(\varepsilon) \in C$ by definition of partitioned Markov chain. $\square$

Pseudo-mixing time starting from the remaining profiles. Consider the distributions μ_i defined above (i.e., the stationary distribution restricted to R_i). We focus here on the convergence time to distributions of the form

$$\nu(\mathbf{y}) = \sum_i \alpha_i \mu_i(\mathbf{y}),$$

for $\alpha_i \geq 0$ and $\sum_i \alpha_i = 1$. Specifically, for every state $\mathbf{x} \in N$, we define the distribution

$$\nu_{\mathbf{x}}(\mathbf{y}) = \sum_i \mu_i(\mathbf{y}) \cdot \mathbf{P}_{\mathbf{x}}\left(X_{\tau_{S\setminus N}} \in T_i \mid \tau_{S\setminus N} \leq \mathcal{T}^{\varepsilon}_{S\setminus N}(\mathbf{x})\right), \tag{3}$$

where $\mathcal{T}^{\varepsilon}_{S\setminus N}(\mathbf{x})$ is the first time step t in which $\mathbf{P}_{\mathbf{x}}(\tau_{S\setminus L} > t) \leq \varepsilon$. Observe that by definition of $\tau_{S\setminus N}$, since the T_i's and N are a partition of S, $X_{\tau_{S\setminus N}} \in \cup_i T_i$ is a certain event for all values of $\tau_{S\setminus N}$. Moreover, by the definition of $\mathcal{T}^{\varepsilon}_{S\setminus N}(\mathbf{x})$, the event $\tau_{S\setminus N} \leq \mathcal{T}^{\varepsilon}_{S\setminus N}(\mathbf{x})$ has non-zero probability and thus we can condition on it. Thus, $\sum_i \mathbf{P}_{\mathbf{x}}\left(X_{\tau_{S\setminus N}} \in T_i \mid \tau_{S\setminus N} \leq \mathcal{T}^{\varepsilon}_{S\setminus N}(\mathbf{x})\right) = 1$. The above is then a valid definition of the α_i's.

Then, we prove the following proposition.

Proposition 3.3. *Let $\mathbf{M}$ be a sequence of n-chains and fix $\varepsilon > 0$. If $\mathbf{M}$ is partitioned for C, then there is a function $\mathcal{T} = \mathcal{T}_{\varepsilon} \notin C$ and a function $p_\star \in C$ such that, for every n and for each $\mathbf{x} \in N$ the corresponding distribution $\nu_{\mathbf{x}}$ is $(\varepsilon, \mathcal{T}(n))$-metastable and the pseudo-mixing time of $\nu_{\mathbf{x}}$ from the state $\mathbf{x}$ is $t^{\{\mathbf{x}\}}_{\nu_{\mathbf{x}}}(4\varepsilon) = O(p_\star(n))$.*

Proof. Fix n. Notice that, the distribution $\nu_{\mathbf{x}}$ is a convex combination of distributions that are metastable for time $\mathcal{T}_i \notin C$: thus, from Lemma 3.1, there exist a function $\mathcal{T} \in C$ such that each such $\nu_{\mathbf{x}}$ is $(\varepsilon, \mathcal{T}(n))$-metastable.

Moreover, from the definition of partitioned chain, we have that $\mathcal{T}^{\varepsilon}_{S\setminus N}(\mathbf{x}) \leq \rho_\star(n) \in C$ for every $\mathbf{x} \in N$. Consider then the function $p_\star(\cdot)$ such that $p_\star(n) = \rho_\star(n) + \max_i t^{T_i}_{\mu_i}(2\varepsilon)$. From Proposition 3.2 and the closure properties of C, it turns out that $p_\star(\cdot) \in C$. We complete the proof by showing that, for every sufficiently large n and every $\mathbf{x} \in N$, $p_\star(n)$ upper bounds the pseudo-mixing time $t^{\{\mathbf{x}\}}_{\nu_{\mathbf{x}}}(4\varepsilon)$ to $\nu_{\mathbf{x}}$ from the state $\mathbf{x}$.

We set $t^\star = p^\star(n)$, and denote with E the event "$\tau_{S\setminus N} \leq \mathcal{T}^{\varepsilon}_{S\setminus N}(\mathbf{x})$" and with $\overline{E}$ its complement. Moreover, we will use $\tau^\star$ as a shorthand for $\tau_{S\setminus N}$.

Recall that X_t denotes the state of the Markov chain at step t and observe that

$$\begin{aligned}
\left\| P^{t^\star}(\mathbf{x}, \cdot) - \nu_{\mathbf{x}} \right\|_{\mathrm{TV}} &= \max_{A \subset S} \left| \mathbf{P}_{\mathbf{x}}\left(X_{t^\star} \in A\right) - \nu_{\mathbf{x}}(A) \right| \\
&= \max_{A \subset S} \left| \mathbf{P}_{\mathbf{x}}\left(X_{t^\star} \in A \wedge E\right) - \nu_{\mathbf{x}}(A) + \mathbf{P}_{\mathbf{x}}\left(X_{t^\star} \in A \wedge \overline{E}\right) \right| \\
&= \max_{A \subset S} \left| \mathbf{P}_{\mathbf{x}}\left(X_{t^\star} \in A \mid E\right)\left(1 - \mathbf{P}_{\mathbf{x}}\left(\overline{E}\right)\right) - \nu_{\mathbf{x}}(A) + \mathbf{P}_{\mathbf{x}}\left(X_{t^\star} \in A \mid \overline{E}\right)\mathbf{P}_{\mathbf{x}}\left(\overline{E}\right) \right| \\
&\leq \max_{A \subset S} \left| \mathbf{P}_{\mathbf{x}}\left(X_{t^\star} \in A \mid E\right) - \nu_{\mathbf{x}}(A) \right| + \mathbf{P}_{\mathbf{x}}\left(\overline{E}\right) \\
&\leq \left\| \mathbf{P}_{\mathbf{x}}\left(X_{t^\star} \mid E\right) - \nu_{\mathbf{x}} \right\|_{\mathrm{TV}} + \varepsilon,
\end{aligned}$$

where the definition of $\mathcal{T}^{\varepsilon}_{S\setminus N}(\mathbf{x})$ implies that $\mathbf{P}_{\mathbf{x}}(E) \geq 1 - \varepsilon > 0$ and then yields the third equality and last inequality. The penultimate inequality, instead, simply follows from the subadditivity of the absolute value and the fact that the difference between two probabilities is upper bounded by 1. As every μ_i is metastable for time $\mathcal{T}_i \notin C$, we have

$$\left\| \mathbf{P}_{\mathbf{x}}\left(X_{t^\star} \mid E\right) - \nu_{\mathbf{x}} \right\|_{\mathrm{TV}} = \left\| \sum_i \sum_{\mathbf{y} \in T_i} \mathbf{P}_{\mathbf{x}}\left(X_{\tau^\star} = \mathbf{y} \mid E\right) \cdot \mathbf{P}_{\mathbf{x}}\left(X_{t^\star} \mid X_{\tau^\star} = \mathbf{y} \wedge E\right) - \nu_{\mathbf{x}} \right\|_{\mathrm{TV}}$$

$$\leq \left\| \sum_i \sum_{\mathbf{y} \in T_i} \mathbf{P}_{\mathbf{x}} \left(X_{\tau^\star} = \mathbf{y} \mid E \right) \left(P^{t^\star - \tau^\star}(\mathbf{y}, \cdot) - \mu_i \right) \right\|_{\mathrm{TV}}$$
$$\leq \sum_i \sum_{\mathbf{y} \in T_i} \mathbf{P}_{\mathbf{x}} \left(X_{\tau^\star} = \mathbf{y} \mid E \right) \left\| P^{t^\star - \tau^\star}(\mathbf{y}, \cdot) - \mu_i \right\|_{\mathrm{TV}} \leq 3\varepsilon,$$

where the definition of $\tau^\star$ yields $X_{\tau^\star} \in T_i$, for some i, which in turns yields the first equality by the law of total probability. In the first inequality above, instead, we use the definition of $\nu_{\mathbf{x}}$ and the fact that by definition of $t^\star$, E implies $t^\star - \tau^\star \geq t^\star - \mathcal{T}^{\varepsilon}_{S \setminus N}(\mathbf{x}) \geq \max_i t^{T_i}_{\mu_i}(2\varepsilon) \geq 0$; the second inequality follows from the subadditivity of the total variation distance; and the last inequality follows from Lemma 2.3 (note that $t^\star - \tau^\star$ satisfies the hypothesis of the lemma: the lower bound is showed above, while the upper bound follows from the fact that the μ_i's are metastable for a time $\mathcal{T}_i \notin C$). Hence, we have for every sufficiently large n and every $\mathbf{x} \in N$, $t^{\{\mathbf{x}\}}_{\nu_{\mathbf{x}}}(4\varepsilon) \leq t^\star = p_\star(n)$. □

4 Logit Dynamics for Potential Games is Asymptotically Metastable

In Section 3 we characterized the (sequences of) Markov chains for which one can prove asymptotic metastability. Here we focus on the specific setting within which metastability of Markov chains has been defined, namely the behavior of the logit dynamics for potential games. We will use the previous characterization to prove that for every potential game and for every starting profile of the game the logit dynamics converges in polynomial time to a distribution that is metastable for a superpolynomial number of steps.

We begin by formally introducing the concepts of potentiality games and logit dynamics.

4.1 Definitions

Potential Games. A *strategic game* $\mathcal{G}$ is a triple $([n], (S_1, \ldots, S_n), \mathcal{U})$, where $[n] = \{1, \ldots, n\}$ is a finite set of players, $(S_1, \ldots, S_n)$ is a family of non-empty finite sets (S_i is the set of strategies available to player i), and $\mathcal{U} = (u_1, \ldots, u_n)$ is a family of utility functions (or payoffs), where $u_i \colon S \to \mathbb{R}$, $S = S_1 \times \ldots \times S_n$ being the set of all strategy profiles, is the utility function of player i. We focus on (exact) *potential games*, i.e., games for which there exists a function $\Phi \colon S \to \mathbb{R}$ such that for every pair of $\mathbf{x}, \mathbf{y} \in S$, with $\mathbf{y} = (\mathbf{x}_{-i}, y_i)$, we have:

$$\Phi(\mathbf{x}) - \Phi(\mathbf{y}) = u_i(\mathbf{y}) - u_i(\mathbf{x}).\ ^{6}$$

Note that we use the standard game theoretic notation $(\mathbf{x}_{-i}, s)$ to mean the vector obtained from $\mathbf{x}$ by replacing the i-th entry with s; i.e. $(\mathbf{x}_{-i}, s) = (x_1, \ldots, x_{i-1}, s, x_{i+1}, \ldots, x_n)$. A strategy profile $\mathbf{x}$ is a Nash equilibrium[7] if, for all i, $u_i(\mathbf{x}) \geq u_i(\mathbf{x}_{-i}, s_i)$, for all $s_i \in S_i$. It is fairly easy to see that local minima of the potential function correspond to the Nash equilibria of the game.

For two vectors $\mathbf{x}, \mathbf{y}$, we denote with $H(\mathbf{x}, \mathbf{y}) = |\{i \colon x_i \neq y_i\}|$ the Hamming distance between $\mathbf{x}$ and $\mathbf{y}$. For every $\mathbf{x} \in S$, $N(\mathbf{x}) = \{\mathbf{y} \in S \colon H(\mathbf{x}, \mathbf{y}) = 1\}$ denotes the set of neighbors of $\mathbf{x}$ and $N_i(\mathbf{x}) = \{\mathbf{y} \in N(\mathbf{x}) \colon \mathbf{y}_{-i} = \mathbf{x}_{-i}\}$ is the set of those neighbors that differ exactly in the i-th coordinate.

Logit dynamics. The logit dynamics has been introduced in [7] and runs as follows: at every time step (i) Select one player $i \in [n]$ uniformly at random; (ii) Update the strategy of player i according to the

[6] Note that this definition is slightly different from the standard one by Monderer and Shapley [28], in which it is required that $\Phi(\mathbf{x}) - \Phi(\mathbf{y}) = u_i(\mathbf{x}) - u_i(\mathbf{y})$. However, it is immediate to see that the two definitions are equivalent. Our definition has been chosen because it helps drawing the connection between logit dynamics and similar concepts in Physics.

[7] In this paper, we only focus on pure Nash equilibria. We avoid explicitly mentioning it throughout.

Boltzmann distribution with parameter β over the set S_i of her strategies. That is, a strategy $s_i \in S_i$ will be selected with probability

$$\sigma_i(s_i \mid \mathbf{x}_{-i}) = \frac{1}{Z_i(\mathbf{x}_{-i})} e^{\beta u_i(\mathbf{x}_{-i}, s_i)}, \tag{4}$$

where $\mathbf{x}_{-i}$ denotes the profile of strategies played at the current time step by players different from i, $Z_i(\mathbf{x}_{-i}) = \sum_{z_i \in S_i} e^{\beta u_i(\mathbf{x}_{-i}, z_i)}$ is the normalizing factor, and $\beta \geq 0$. One can see parameter β as the inverse of the noise or, equivalently, the *rationality level* of the system: indeed, from (4), it is easy to see that for $\beta = 0$ player i selects her strategy uniformly at random, for $\beta > 0$ the probability is biased toward strategies promising higher payoffs, and for β that goes to infinity player i chooses her best response strategy (if more than one best response is available, she chooses one of them uniformly at random).

The above dynamics defines a Markov chain $\{X_t\}_{t \in \mathbb{N}}$ with the set of strategy profiles as state space, and where the transition probability $P(\mathbf{x}, \mathbf{y})$ from profile $\mathbf{x} = (x_1, \ldots, x_n)$ to profile $\mathbf{y} = (y_1, \ldots, y_n)$ is zero if $H(\mathbf{x}, \mathbf{y}) \geq 2$ and it is $\frac{1}{n}\sigma_i(y_i \mid \mathbf{x}_{-i})$ if the two profiles differ exactly at player i. More formally, we can define the logit dynamics as follows.

Definition 4.1. Let $\mathcal{G} = ([n], (S_1, \ldots, S_n), \mathcal{U})$ be a strategic game and let $\beta \geq 0$. The *logit dynamics* for $\mathcal{G}$ is the Markov chain $\mathcal{M}_\beta = (S, P)$ where $S = S_1 \times \cdots \times S_n$ and

$$P(\mathbf{x}, \mathbf{y}) = \frac{1}{n} \cdot \begin{cases} \sigma_i(y_i \mid \mathbf{x}_{-i}), & \text{if } \mathbf{y}_{-i} = \mathbf{x}_{-i} \text{ and } y_i \neq x_i; \\ \sum_{i=1}^{n} \sigma_i(y_i \mid \mathbf{x}_{-i}), & \text{if } \mathbf{y} = \mathbf{x}; \\ 0, & \text{otherwise;} \end{cases} \tag{5}$$

where $\sigma_i(y_i \mid \mathbf{x}_{-i})$ is defined in (4).

The Markov chain defined by (5) is ergodic [7]. Hence, the logit dynamics will eventually converge to a stationary distribution π, such that $\pi P = \pi$. As in [4], we call the stationary distribution π of the Markov chain defined by the logit dynamics on a game $\mathcal{G}$, the *logit equilibrium* of $\mathcal{G}$.

For the class of potential games the stationary distribution of the logit dynamics is the well-known *Gibbs measure*.

Theorem 4.1 ([7])**.** *If $\mathcal{G} = ([n], (S_1, \ldots, S_n), \mathcal{U})$ is a potential game with potential function Φ, then the Markov chain given by* (5) *is reversible with respect to the Gibbs measure $\pi(\mathbf{x}) = \frac{1}{Z} e^{-\beta\Phi(\mathbf{x})}$, where $Z = \sum_{\mathbf{y} \in S} e^{-\beta\Phi(\mathbf{y})}$ is the normalizing constant.*

It is worthwhile to notice that logit dynamics for potential games and Glauber dynamics for Gibbs distributions are two ways of looking at the same Markov chain (see [7] for details). This, in particular, implies that we can write

$$\sigma_i(s_i \mid \mathbf{x}_{-i}) = \frac{e^{-\beta\Phi(\mathbf{x}_{-i}, s_i)}}{\sum_{z \in S_i} e^{-\beta\Phi(\mathbf{x}_{-i}, z)}}.$$

4.2 All Asymptotically Well-Defined Potential Games Are Asymptotically Metastable

Here, we are interested in showing that mestastable distributions are reached in polynomial time and remain stable for a superpolynomial number of steps. In other words, we assume that the class C of functions for which we will prove asymptotic metastability of the Markov chains describing the logit dynamics consists of the class of all functions that are at most polynomial. Since C is well-understood in this setting, we will omit it in the rest of this section. Moreover, when the logit dynamics for a game is (not) asymptotically metastable, we say for brevity that the game itself is (not) asymptotically metastable. Similarly, when the Markov chains describing the logit dynamics run on a game is (not)

partitioned, we will say for brevity that the game is (not) partitioned by the logit dynamics or that the logit dynamics is (not) partitioned.

Before our analysis, we need to highlight some assumptions that we make henceforth. First, bounds on the convergence of the logit dynamics are usually measured on the number of players of the game on which the dynamics is run. This is in contrast with the previous section, where we discussed that convergence of Markov chains are instead measured in the logarithm of the number of states. To address this issue, we will assume that the logarithm of the maximal number of strategies available to a player is at most a polynomial in n. Specifically, given a sequence of n-player games $\mathbf{G}$, one for each number n of players, we denote as $m(\cdot)$ the function such that $m(n)$ is the maximum number of strategies available to a player in $\mathcal{G}_n$, where $\mathcal{G}_n$ is the game in the sequence $\mathbf{G}$ with exactly n-players. Then, we will assume that the function $\log m(\cdot)$ is at most polynomial in its input. We can easily drop this assumption by asking for results that are asymptotic in $\log |S|$, where $|S|$ denotes the function returning the number of profiles of the game: each one of our proofs can be rewritten according to this requirement with very small changes. Note that since $|S|$ for a game with n players is at most $m(n)^n$, this requirement is equivalent to asking for results asymptotic in n and in the logarithm of the function m.

Note also that we focus only on n-player games and values of β such that the mixing time of the logit dynamics is at least super-polynomial in n, otherwise, as indicated above, the stationary distribution enjoys the desired properties of stability and convergence. Throughout the rest of the paper we will denote with β_0 the smallest value of β such that the mixing time is not polynomial.

Finally, as discussed in the introduction, we are aiming to give results asymptotic in the number n of players about the behavior of logit dynamics run on potential games. Clearly, it makes sense to give asymptotic results about the property of an object, only if this property is asymptotically well-defined, that is, the object is uniquely defined for infinitely many values of the parameter according to which we compute the asymptotic and the property of this object does not depend "chaotically" on this parameter. Here the object of interest are games (specifically, potential games) and the property of interest is the behavior of the logit dynamics for these games and a given rationality level β. Thus, in our setting, it makes sense to give asymptotic results only when a potential game is uniquely defined for infinitely many n and the behavior of the logit dynamics for this potential game and a given β is not chaotic as n increases. We call these games *asymptotically well-behaved* for the given β. Specifically, we have the following definition.

Definition 4.2. A sequence $\mathbf{G}$ of n-player potential games is *asymptotically well-behaved* for β, if there is n_0, a polynomial p and a super-polynomial q such that for every $n \geq n_0$ and for every $L \subseteq \mathcal{S}_n$, either $B(L) \geq \frac{1}{p(n)}$ or $B(L) \leq \frac{1}{q(n)}$, with $B(L)$ being the bottleneck ratio of L in the the Markov chain defined by the logit dynamics for game $\mathbf{G}$ and rationality level β. Henceforth, we will say that the functions p, q are *generated* by the sequence of games $\mathbf{G}$.

Intuitively, these are the only games for which it makes sense to give asymptotic results, since if there is a subset of profiles for which the bottleneck ratio is neither at least the inverse of a polynomial nor at most the inverse of a super-polynomial, then, as we formally prove below in Lemmas 4.6 and 4.7, we cannot say if this subset is "easy-to-leave" or "hard-to-leave". Hence, depending on the distribution of interest, no asymptotic result can be given when the dynamics starts from these profiles.[8] In Appendix B,

[8]It turns out that for the metastable distributions we identify in Proposition 3.1 we might be able to give asymptotic results even when not all the sets of profiles have their bottleneck ratio classified as either polynomial or super-polynomial. The intuition is the following: consider a large subset L, not classified, of a subset A classified as "hard-to-leave". The stationary distribution restricted to A might give a large weight to the profiles in L and therefore the unknown behavior of L might not be an obstacle to proving that in A it is "easy-to-mix". Thus we can asymptotically describe the behavior of the logit dynamics starting from L, even if we do not know how easy it is to leave L. We show in Appendix C that there are simple games that are asymptotically metastable, but that are not captured by Definition 4.2, and give a technical definition to describe these games. Nevertheless, we see this more as a technical nuisance than a conceptual contribution. We therefore prefer to keep our focus on the clean, simple and intuitive definition of asymptotically well-behaved games.

we indeed show several examples of these "bad" games for which asymptotic results about metastability do not make sense, and that our definition must necessarily rule out. We also show how these examples guide us to Definition 4.2.

We are now ready to formally state the main result of this section, namely that every asymptotically well-behaved potential game is asymptotic metastable. More specifically, we will prove the following theorem.

Theorem 4.2. *Let $\mathbf{G}$ be a sequence of n-player potential games and let Δ be the function that, for every n, gives the Lipschitz constant of the potential function Φ_n of $\mathcal{G}_n$, i.e.,*

$$\Delta(n) := \max\left\{\Phi_n(\mathbf{x}) - \Phi_n(\mathbf{y})\colon H(\mathbf{x},\mathbf{y}) = 1\right\}.$$

Then, for every function ρ at most polynomial, if $\mathbf{G}$ is asymptotically well-behaved for $\beta_0 \leq \beta \leq \frac{\rho(n)}{\Delta(n)}$, then it is asymptotically metastable for β.

The dependence on $\Delta(n)$ is a by-product of the fact that the logit dynamics is not invariant to scaling of the utility function, i.e., scaling the utility function of a certain factor requires to inversely scale β to get the same logit dynamics (see [4] for a discussion). In a sense, $\beta\Delta(n)$ is the natural parameter that describes the logit dynamics. Then, from this point of view, the requirement on β in the above theorem becomes almost natural: we, indeed, require that $\beta\Delta(n)$ is sufficiently large in order for the mixing time to be not polynomial, but we also require that $\beta\Delta(n)$ is a polynomial. This assumption is in general necessary because when $\beta\Delta(n)$ is high enough logit dynamics roughly behaves as best-response dynamics. Moreover, in this case, the only metastable distributions have to be concentrated around the set of Nash equilibria. This is because for $\beta\Delta(n)$ very high, it is extremely unlikely that a player leaves a Nash equilibrium. Then, the hardness results about the convergence of best-response dynamics for potential games, cf. e.g. [17], imply that the convergence to metastable distributions for high $\beta\Delta(n)$ is similarly computationally hard.

The proof builds upon Theorem 3.1 and proves that the Markov chain corresponding to the logit dynamics for an asymptotically well-behaved potential game is partitioned, whenever β is as in the statement.

4.2.1 Proof of Theorem 4.2

In order to prove Theorem 4.2, we introduce a (computationally infeasible) algorithm that, given a sequence of n-player games $\mathbf{G}$ and n, computes subsets $R_1, \ldots, R_k$ of the set S of profiles of $\mathcal{G}_n$ and a partition $T_1, \ldots, T_k, N$ of S. Next we show that under the condition that the potential game is asymptotically well-behaved, the sets returned by this algorithm enjoy the properties required by the definition of partitioned Markov chain. The procedure works its way by finding subsets of profiles that act as super-polynomial bottlenecks for the Markov chain. The algorithm $\mathcal{A}_{p,q}$ takes in input an asymptotically well-behaved sequence of n-player potential games $\mathbf{G}$, a rationality level β, a constant $\varepsilon > 0$ and n; it is parametrized by two functions p and q generated by $\mathbf{G}$.

Algorithm 4.1 ($\mathcal{A}_{p,q}$)**.** Set $N = S$ and $i = 1$. While there is a set $L \subseteq N$ with $\pi(L) \leq 1/2$ such that $B(L) \leq 1/q(n)$, do:

1. Denote with R_i one such subset with the smallest stationary probability;
2. Denote with T_i the largest subset of R_i such that for every $\mathbf{y} \in T_i$,

$$\mathbf{P}_{\mathbf{y}}\left(\tau_{S\setminus R_i} \leq t_{\text{mix}}^{R_i}(\varepsilon)\right) \leq \varepsilon;$$

3. If T_i is not empty, return R_i and T_i, delete from N all profiles contained in T_i and increase i. Otherwise, terminate the algorithm.

Observe that if there is a disconnected set L such that $B(L) \leq 1/q(n)$, then each connected component L' of L will have $B(L') \leq 1/q(n)$ and smaller stationary probability: hence, the set R_i returned by the algorithm will be connected. Note also that by Theorem 2.2 and the assumption that we are considering only cases in which the mixing time is super-polynomial, the algorithm above enters at least once in the loop (and thus at least a subset R_i is computed).

Clearly the sets R_i returned by the algorithm enjoy the property of having super-polynomial bottleneck ratio and the sets T_i satisfy the requirement that, starting from any $\mathbf{x} \in T_i$, it is unlikely to leave R_i quickly. It is then left to prove that the mixing time of the chains restricted to R_i is polynomial and that it is easy to leave the set N. This follows from the following propositions that are proved in the next sections.

Proposition 4.1. *Let $\beta_0 \leq \beta \leq \frac{\rho(n)}{\Delta(n)}$, with ρ at most polynomial in its input. Let $\mathbf{G}$ be a sequence of n-player potential games asymptotically well-behaved for β. Let p, q be the functions generated by $\mathbf{G}$. Consider the sequence of sets R_i returned by $\mathcal{A}_{p,q}$. Then, for every $\varepsilon > 0$, $t_{\text{mix}}^{R_i}(\varepsilon)$ is at most polynomial.*

Proposition 4.2. *Let $\beta_0 \leq \beta \leq \frac{\rho(n)}{\Delta(n)}$, with ρ at most polynomial in its input. Let $\mathbf{G}$ be a sequence of n-player potential games asymptotically well-behaved for β. Let p, q be the functions generated by $\mathbf{G}$. Consider the set N returned by $\mathcal{A}_{p,q}$. Then, for every $\varepsilon > 0$, $\mathcal{T}_{S\setminus N}^{\varepsilon}(\mathbf{x})$ is at most polynomial, where, for $\mathbf{x} \in N$, $\mathcal{T}_{S\setminus N}^{\varepsilon}(\mathbf{x})$ is defined as the first time step t in which $\mathbf{P}_{\mathbf{x}}(\tau_{S\setminus N} > t) \leq \varepsilon$.*

This proves that asymptotically well-behaved potential games are partitioned by the logit dynamics and, hence, they are asymptotically metastable, concluding the proof of Theorem 4.2.

4.2.2 Proof of Proposition 4.1

A high level idea of the proof of Proposition 4.1 is discussed next. We first give a spectral characterization of the transition matrix defined in (2) (Lemma 4.2). Then we show that no subset of R_i has small bottleneck ratio in the restricted chain (Lemma 4.3). Note that this does not directly follow from R_i being the smallest subset with super-polynomial bottleneck ratio. Indeed, the bottleneck ratio of a subset depends on the dynamics according to which it is computed. Thus, a subset can have a small bottleneck ratio when computed within the reference frame of the restricted dynamics, but not when we refer to the original dynamics. Nevertheless, we show that this is not the case. Specifically, we will show that for asymptotically well-behaved games there is a relationship between the bottleneck ratio of a subset of profiles in the restricted and in the original dynamics. Finally, the result follows from the known relationship among mixing time, relaxation time and bottleneck ratio (see Theorems 2.1 and 2.3).

Spectral property of logit dynamics restrictions

In [3] it has been shown that all the eigenvalues of the transition matrix of logit dynamics for potential games are non-negative. The technique used in that proof can be generalized to work also for some restrictions of these matrices.

To begin, we note that the definition of reversibility can be extended in a natural way to every square matrix and probability distribution over the set of rows of the matrix. We then state a fairly standard result relating eigenvalues of matrices to certain inner products.

Lemma 4.1. *Let P be a square matrix on state space S and π be a probability distribution on S. If P is reversible with respect to π and has no negative eigenvalues then for every function $f : S \to \mathbb{R}$ we have*

$$\langle Pf, f \rangle_\pi := \sum_{\mathbf{x} \in S} \pi(\mathbf{x})(Pf)(\mathbf{x}) f(\mathbf{x}) \geq 0.$$

Proof. Let $\lambda_1, \ldots, \lambda_s$, $s = |S|$, be the eigenvalues of P. Moreover, let $f_1, \ldots, f_s$ denote their corresponding eigenfunctions. For every $\mathbf{x} \in S$, we then have $(Pf_i)(\mathbf{x})f_i(\mathbf{x}) = \lambda_i f_i(\mathbf{x})$. Since P is reversible then we know that the eigenfunctions assume real values and that they form an orthonormal basis for the space $(\mathbb{R}^s, \langle \cdot, \cdot \rangle_\pi)$ (see, e.g., Lemma 12.2 in [25]). Then every real-valued function f defined upon S can be expressed as a linear combination of the f_i's. Thus, there exist α_i's in $\mathbb{R}$ such that

$$\sum_{\mathbf{x}\in S} \pi(\mathbf{x})(Pf)(\mathbf{x})f(\mathbf{x}) = \sum_{\mathbf{x}\in S} \pi(\mathbf{x}) \sum_{i=1}^{s} \alpha_i^2 (Pf_i)(\mathbf{x})f_i(\mathbf{x}) = \sum_{\mathbf{x}\in S} \pi(\mathbf{x}) \sum_{i=1}^{s} \alpha_i^2 \lambda_i f_i^2(\mathbf{x}) \geq 0. \qquad \square$$

To specify the restrictions of the transition matrix we are interested in, let $\mathcal{G}$ be a game with profile space S and let P be the transition matrix of the logit dynamics for $\mathcal{G}$; we say that a $|A| \times |A|$ matrix P', with $A \subseteq S$, is a *nice restriction* of P if there exists $L \subseteq A$, $L \neq \emptyset$, such that $P'(\mathbf{x}, \mathbf{x}) \geq P(\mathbf{x}, \mathbf{x})$ for $\mathbf{x} \in L$, $P'(\mathbf{x}, \mathbf{y}) = P(\mathbf{x}, \mathbf{y})$ if $\mathbf{x}, \mathbf{y} \in L$, $\mathbf{x} \neq \mathbf{y}$, and is 0 otherwise. Note that P is a nice restriction of itself. We generalize the result given in [3] to nice restrictions of the transition matrix of logit dynamics for potential games.

Lemma 4.2. *Let $\mathcal{G}$ be a game with profile space S, let P be the transition matrix of the logit dynamics for $\mathcal{G}$ and let P' be a nice restriction of P with state space A. If P is reversible then no eigenvalue of P' is negative.*

Proof. Firstly, note that since P is reversible with respect to π then the nice restriction P', defined upon a subset of states A, is reversible with respect to π' defined as π restricted to A, i.e., $\pi'(\mathbf{x}) = \pi(\mathbf{x})/\pi(A)$ for $\mathbf{x} \in A$.

Assume for sake of contradiction that there exists an eigenvalue $\lambda < 0$ of P'. Let f_λ be an eigenfunction of λ. Note that since P' is reversible then f_λ is real-valued. By definition, $f_\lambda \neq \mathbf{0}$; hence, since $\lambda < 0$ and as $(P'f_\lambda)(\mathbf{x}) = \lambda f_\lambda(\mathbf{x})$, then for every profile $\mathbf{x} \in A$ such that $f_\lambda(\mathbf{x}) \neq 0$ we have $\mathsf{sign}\left((P'f_\lambda)(\mathbf{x})\right) \neq \mathsf{sign}\left(f_\lambda(\mathbf{x})\right)$ and thus

$$\langle P'f_\lambda, f_\lambda \rangle_{\pi'} = \sum_{\mathbf{x}\in A} \pi'(\mathbf{x})(P'f_\lambda)(\mathbf{x})f_\lambda(\mathbf{x}) < 0.$$

Let L denote the maximal subset of A for which P' is a nice restriction of P. Let us denote with P^L the transition matrix on the state space A such that $P^L(\mathbf{x}, \mathbf{y}) = P(\mathbf{x}, \mathbf{y})$ for every $\mathbf{x}, \mathbf{y} \in L$ and $P^L(\mathbf{x}, \mathbf{y}) = 0$ otherwise. Then we can write P' as $P^L + (P' - P^L)$: by the definition of nice restriction $(P' - P^L)$ is a non-negative diagonal matrix. Therefore, $(P' - P^L)$ is reversible with respect to π'. Since the eigenvalues of a diagonal matrix are exactly the diagonal elements, we have that $(P' - P^L)$ has non-negative eigenvalues and then, by Lemma 4.1, $\langle (P' - P^L)f_\lambda, f_\lambda \rangle_{\pi'} \geq 0$. Moreover, for every i and for every $\mathbf{z}_{-i}$, we denote with $P_{i,\mathbf{z}_{-i}}$ the matrix such that for every $\mathbf{x}, \mathbf{y} \in A$

$$P_{i,\mathbf{z}_{-i}}(\mathbf{x}, \mathbf{y}) = \frac{1}{nZ_i(\mathbf{z}_{-i})} \begin{cases} e^{\beta u_i(\mathbf{y})}, & \text{if } \mathbf{x}_{-i} = \mathbf{y}_{-i} = \mathbf{z}_{-i} \text{ and } \mathbf{x}, \mathbf{y} \in L; \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

Observe that $P_{i,\mathbf{z}_{-i}}$ has at least one non-zero row and that all non-zero rows of $P_{i,\mathbf{z}_{-i}}$ are the same. Thus $P_{i,\mathbf{z}_{-i}}$ has rank 1, and hence since it is a non-negative matrix all its eigenvalues are non-negative [22][9]. Moreover, since all off-diagonal entries of $P_{i,\mathbf{z}_{-i}}$ are either 0 or equal to the corresponding entry of P' we can conclude that $P_{i,\mathbf{z}_{-i}}$ is reversible with respect to π'. Thus, Lemma 4.1 yields $\langle P_{i,\mathbf{z}_{-i}} f_\lambda, f_\lambda \rangle_{\pi'} \geq 0$. Finally, observe that $P^L = \sum_i \sum_{\mathbf{z}_{-i}} P_{i,\mathbf{z}_{-i}}$. Hence from the linearity of the inner product, it follows that $\langle P'f_\lambda, f_\lambda \rangle_{\pi'} \geq 0$ and thus we reach a contradiction. $\square$

It is immediate to see that the restricted chain $\mathring{P}_L$ defined in (2) is a nice restriction of P and hence all its eigenvalues are non-negative by the lemma above.

[9] This result about the eigenvalues of matrices with rank 1 appears as an exercise at page 61 of [22] and in [32].

Mixing time of the restricted chains

Before bounding the mixing time of the restricted chain we prove a very important preliminary lemma.

Lemma 4.3. *Let $\beta \geq \beta_0, \varepsilon > 0$, and let $\mathbf{G}$ be a sequence of n-player potential games asymptotically well-behaved for β. Let p, q the functions generated by $\mathbf{G}$. Consider the sequence of sets R_i returned by $\mathcal{A}_{p,q}$. Then, for n sufficiently large and for every $A \subseteq R_i$, we have*

$$B_{R_i}(A) \geq \frac{1}{p(n)} - \frac{1}{\ell(n)},$$

where ℓ is at least super-polynomial.

Proof. Let us postpone the exact definition of ℓ and suppose, by contradiction, that there are infinitely many n for which there is $A \subset R_i$ such that $B_{R_i}(A) < \frac{1}{p(n)} - \frac{1}{\ell(n)}$.

We will show that for n sufficiently large either $B(A) \leq 1/q(n)$ or $B(\overline{A}) \leq 1/q(n)$, where $\overline{A} = R_i \setminus A$. Then, since they are contained in R_i and hence their stationary probability is less than $\pi(R_i)$, one of these set must be chosen before R_i by $\mathcal{A}_{p,q}$. But since in the third step of Algorithm 4.1 either at least one element of such sets should be deleted from N or the algorithm terminates, as a consequence, we have that R_i cannot be returned by the algorithm, thus a contradiction.

Consider the function $v(\cdot)$ that sets $v(n) = \frac{\pi(A)}{Q(A, S \setminus R_i)}$. We distinguish two cases depending on how v evolves as n grows.

If $v(\cdot)$ is at least super-polynomial in the input: We have

$$\begin{aligned} B(A) &= \frac{Q(A, S \setminus A)}{\pi(A)} = \frac{Q(A, R_i \setminus A)}{\pi(A)} + \frac{Q(A, S \setminus R_i)}{\pi(A)} \\ &= \frac{\sum_{\mathbf{x} \in A} \sum_{\mathbf{y} \in R_i \setminus A} \pi(\mathbf{x}) P(\mathbf{x}, \mathbf{y})}{\pi(A)} + \frac{Q(A, S \setminus R_i)}{\pi(A)} \\ &= \frac{\sum_{\mathbf{x} \in A} \sum_{\mathbf{y} \in R_i \setminus A} \pi_{R_i}(\mathbf{x}) \mathring{P}_{R_i}(\mathbf{x}, \mathbf{y})}{\pi_{R_i}(A)} + \frac{Q(A, S \setminus R_i)}{\pi(A)} \\ &= B_{R_i}(A) + \frac{Q(A, S \setminus R_i)}{\pi(A)} < \frac{1}{p(n)} + \frac{1}{v(n)} - \frac{1}{\ell(n)}. \end{aligned}$$

By taking $\ell(n) \leq v(n)$ for each n sufficiently large, we have that $B(A) < \frac{1}{p(n)}$. Then, since $\mathbf{G}$ is asymptotically well-behaved, from Definition 4.2 it follows that $B(A) \leq \frac{1}{q(n)}$.

If $v(\cdot)$ is polynomial in the input: Note that $\frac{Q(A, S \setminus R_i)}{\pi(R_i)} + \frac{Q(\overline{A}, S \setminus R_i)}{\pi(R_i)} = B(R_i) \leq \frac{1}{q(n)}$, otherwise R_i was not returned by the algorithm. Hence, we obtain

$$Q(A, S \setminus R_i) \leq \frac{1}{q(n)} \cdot \pi(R_i) \qquad \text{and} \qquad Q(\overline{A}, S \setminus R_i) \leq \frac{1}{q(n)} \cdot \pi(R_i).$$

From the first of these inequalities, we have $\pi(A) \leq \frac{v(n)}{q(n)} \cdot \pi(R_i)$. Hence

$$\frac{Q(A, \overline{A})}{\pi(R_i)} \leq \frac{v(n)}{q(n)} \cdot \frac{Q(A, \overline{A})}{\pi(A)} = \frac{v(n)}{q(n)} \cdot B_{R_i}(A) < \frac{v(n)}{q(n)} \left(\frac{1}{p(n)} - \frac{1}{\ell(n)} \right).$$

Then we obtain

$$B(\overline{A}) = \frac{Q(\overline{A}, S \setminus \overline{A})}{\pi(\overline{A})} = \frac{Q(\overline{A}, A)}{\pi(R_i) - \pi(A)} + \frac{Q(\overline{A}, S \setminus R_i)}{\pi(R_i) - \pi(A)}$$

$$
\begin{aligned}
\text{(by reversibility of } P) \quad &= \frac{Q\left(A, \overline{A}\right)}{\pi\left(R_i\right) - \pi\left(A\right)} + \frac{Q\left(\overline{A}, S \setminus R_i\right)}{\pi\left(R_i\right) - \pi\left(A\right)} \\
&\leq \frac{v(n)}{q(n)} \left(\frac{1}{p(n)} - \frac{1}{\ell(n)}\right) \left(1 - \frac{v(n)}{q(n)}\right)^{-1} + \frac{1}{q(n)} \left(1 - \frac{v(v)}{q(n)}\right)^{-1} \\
&= O\left(\frac{1}{q(n) - v(n)}\right),
\end{aligned}
$$

where the upper bounds hold for each choice of super-polynomial function ℓ. Since $q(n) - v(n)$ evolves at least as a super-polynomial, if n is sufficiently large, $B(\overline{A}) < \frac{1}{p(n)}$. Then, since $\mathbf{G}$ is asymptotically well-behaved, from Definition 4.2 it follows that $B\left(\overline{A}\right) \leq \frac{1}{q(n)}$. □

Now we are ready to prove the mixing time of the chain restricted to R_i is polynomial.

Proof of Proposition 4.1. Fix n. Consider the set of profiles $A_\star \subset R_i$ that minimizes $B_{R_i}(A)$ among all $A \subset R_i$ such that $\pi_{R_i}(A) \leq 1/2$. By Lemma 4.3, $B_{R_i}(A_\star) \geq 1/p(n) - 1/\ell(n)$ for each n sufficiently large.

Moreover, for each n and each $\mathbf{x} \in R_i$, since $|S| \leq m(n)^n$, it follows that

$$
\log \frac{1}{\pi_{R_i}(\mathbf{x})} \leq \log \frac{|S| e^{-\beta \Phi_{\min}}}{e^{-\beta \Phi_{\max}}} \leq \log \frac{e^{n \log m(n)} e^{-\beta \Phi_{\min}}}{e^{-\beta \Phi_{\max}}} = n \log m(n) + \beta \left(\Phi_{\max} - \Phi_{\min}\right),
$$

where $\Phi_{\max}$ and $\Phi_{\min}$ denote the maximum and minimum of the potential Φ overall possible strategy profiles. Since $\Phi_{\max} - \Phi_{\min} \leq n \cdot \Delta(n)$ and $\beta \leq \rho(n)/\Delta(n)$, then

$$
\log \frac{1}{\pi_{R_i}(\mathbf{x})} \leq n \cdot \left(\log m(n) + \rho(n)\right).
$$

Then, from Lemma 4.2 and the properties of the relaxation time (see Theorems 2.3 and 2.1) it follows that the mixing time is

$$
t_{\text{mix}}^{R_i}(\varepsilon) \leq \left(\frac{1}{p(n)} - \frac{1}{\ell(n)}\right)^{-2} \cdot \left(n \log m(n) + \rho(n)\right) \cdot 2 \log \frac{4}{\varepsilon} = O(p_\star(n)).
$$

Since p, $\log m$ and ρ are at most polynomial and ℓ is at least a super-polynomial, then $p_\star$ is at most polynomial in its input and the lemma follows. □

4.2.3 Proof of Proposition 4.2

In order to prove Proposition 4.2, we show that there is a strong relationship between hitting time and metastability (see Lemma 4.8 and Lemma 4.9) and, in particular, that high hitting time implies the existence of a subset with small bottleneck ratio. Note that N contains subsets of small bottleneck ratio only if at some iteration T_i is empty. Therefore, it is sufficient to prove that the cores are not empty for asymptotically well-behaved games (see Lemma 4.10).

The relation between bottleneck ratio and hitting time

For a game $\mathcal{G}_n$ with potential function Φ and profile space S, and a rationality level β, let P be the transition matrix of the Markov chain defined by the logit dynamics on $\mathcal{G}_n$. For a non-empty $L \subseteq S$, we denote with $P_{\overline{L}}$ the matrix

$$
P_{\overline{L}}(\mathbf{x}, \mathbf{y}) = \begin{cases} P(\mathbf{x}, \mathbf{y}) & \text{if } \mathbf{x}, \mathbf{y} \in L; \\ 0 & \text{otherwise.} \end{cases} \tag{7}
$$

Let $\lambda_1^{\overline{L}} \geq \lambda_2^{\overline{L}} \geq \ldots \geq \lambda_{|S|}^{\overline{L}}$ be the eigenvalues of $P_{\overline{L}}$: notice that $\lambda_1^{\overline{L}}$ can be different from 1 since the matrix $P_{\overline{L}}$ is not stochastic. Observe that $P_{\overline{L}}$ is a nice restriction of P. Hence, Lemma 4.2 implies that $\lambda_1^{\overline{L}} \geq \lambda_2^{\overline{L}} \geq \ldots \geq \lambda_{|S|}^{\overline{L}} \geq 0$, and thus for $\lambda_{\max}^{\overline{L}}$, the largest eigenvalue of $P_{\overline{L}}$ in absolute value, we have:

$$\lambda_{\max}^{\overline{L}} = \max_i |\lambda_i^{\overline{L}}| = \lambda_1^{\overline{L}}.$$

We start with two characterizations of $1 - \lambda_{\max}^{\overline{L}}$ in terms of bottleneck ratio. The first one is an easy extension of the similar characterization of the spectral gap of stochastic matrices.

Lemma 4.4. *For finite β and every $\emptyset \neq L \subseteq S$, $1 - \lambda_{\max}^{\overline{L}} \leq B(L)$.*

Proof. Define the function $\varphi_L : S \to [0,1]$ to be such that $\varphi_L(\mathbf{x}) = \pi(L)$ if $\mathbf{x} \in L$, and $\varphi_L(\mathbf{x}) = 0$ otherwise. Consider now the function

$$\mathcal{E}_P(\varphi_L) := \frac{1}{2} \sum_{\mathbf{x},\mathbf{y} \in S} \pi(\mathbf{x}) P(\mathbf{x},\mathbf{y}) (\varphi_L(\mathbf{x}) - \varphi_L(\mathbf{y}))^2. \tag{8}$$

By Theorem 4.1, $\pi(L) \neq 0$ and then $\mathbf{E}_\pi\left[\varphi_L^2\right] = \pi(L)^3 \neq 0$. Moreover, by recalling the definition of ∂L as the set of profiles $\mathbf{x} \in L$ that have at least one neighbor profile in $S \setminus L$ and denoting with $E(A_1, A_2)$ the pairs of neighbor profiles $(\mathbf{x}, \mathbf{y})$ such that $\mathbf{x} \in A_1$ and $\mathbf{y} \in A_2$. We have:

$$\begin{aligned}
\mathcal{E}_P(\varphi_L) &= \frac{\pi(L)^2}{2} \left(\sum_{(\mathbf{x},\mathbf{y}) \in E(L, S\setminus L)} \pi(\mathbf{x}) P(\mathbf{x},\mathbf{y}) + \sum_{(\mathbf{x},\mathbf{y}) \in E(S\setminus L, L)} \pi(\mathbf{x}) P(\mathbf{x},\mathbf{y}) \right) \\
&= \pi(L)^2 \sum_{\mathbf{x} \in \partial L} \pi(\mathbf{x}) \sum_{\substack{\mathbf{y} \in S \setminus L: \\ H(\mathbf{x},\mathbf{y}) = 1}} P(\mathbf{x},\mathbf{y}) = \pi(L)^2 Q(L, S \setminus L),
\end{aligned}$$

where we used the reversibility of P in the penultimate equality. Hence, we have $\frac{\mathcal{E}_P(\varphi_L)}{\mathbf{E}_\pi\left[\varphi_L^2\right]} = B(L)$. The lemma follows since $1 - \lambda_{\max}^{\overline{L}} \leq \frac{\mathcal{E}_P(\varphi_L)}{\mathbf{E}_\pi\left[\varphi_L^2\right]}$ (see Lemma A.1 in Appendix). □

The second characterization may be proved in exactly the same way as a similar well-known characterization for the spectral gap of stochastic matrices (see, for example, Section 13.3.3 in [25]).

Lemma 4.5. *For every $\emptyset \neq L \subseteq S$,*

$$1 - \lambda_{\max}^{\overline{L}} \geq \frac{(B_\star^L)^2}{2}.$$

Finally, let us recall a couple of lemmata relating $\tau_{S\setminus L}$ and $\lambda_{\max}^{\overline{L}}$ and already stated in e.g. [29].

Lemma 4.6. *For a reversible Markov chain with state space S, every $L \subseteq S$ and every t it holds that*

$$\max_{\mathbf{x} \in L} \mathbf{P}_{\mathbf{x}}\left(\tau_{S\setminus L} > t\right) \geq \exp\left(t \log \lambda_{\max}^{\overline{L}}\right).$$

Lemma 4.7. *For a reversible Markov chain with state space S, every $L \subseteq S$ and every t it holds that*

$$\mathbf{P}_{\mathbf{x}}\left(\tau_{S\setminus L} > t\right) \leq \exp\left(t \log \lambda_{\max}^{\overline{L}} + \frac{1}{2} \log \frac{1}{\pi_L(\mathbf{x})}\right),$$

where $\pi_L(\mathbf{x})$ has been defined in (1).

Since the statement of Lemma 4.7 is slightly different from the ones found in previous literature, we provide a proof in Appendix A.2 for sake of completeness.

The above lemmata represent the main ingredients to prove the following relations between bottleneck ratio and hitting time.

Lemma 4.8. *Let $\mathcal{G}_n$ be a potential game with profile space S and let P be the transition matrix of the logit dynamics for $\mathcal{G}_n$. Then for finite β and $L \subset S$, $L \neq \emptyset$, we have*

$$\min_{\mathbf{x}\in L} \mathbf{P}_{\mathbf{x}}\left(\tau_{S\setminus L} \leq t\right) \leq t \cdot \frac{B(L)}{1-B(L)}.$$

Proof. We observe:

$$\begin{aligned}
\min_{\mathbf{x}\in L} \mathbf{P}_{\mathbf{x}}\left(\tau_{S\setminus L} \leq t\right) &= 1 - \max_{\mathbf{x}\in L} \mathbf{P}_{\mathbf{x}}\left(\tau_{\overline{L}} > t\right) \\
\text{(by Lemma 4.6)} \qquad &\leq 1 - \exp\left(t \log \lambda_{\max}^{\overline{L}}\right) \\
&= 1 - \exp\left(t \log(1 - (1 - \lambda_{\max}^{\overline{L}}))\right) \\
\text{(since } 1 - a \geq e^{-\frac{a}{1-a}}\text{)} \qquad &\leq 1 - \exp\left(-t \frac{1 - \lambda_{\max}^{\overline{L}}}{\lambda_{\max}^{\overline{L}}}\right) \\
\text{(by Lemma 4.4)} \qquad &\leq 1 - \exp\left(-t \cdot \frac{B(L)}{1-B(L)}\right) \\
\text{(since } 1 - e^{-a} \leq a\text{)} \qquad &\leq t \cdot \frac{B(L)}{1-B(L)}.
\end{aligned}$$

□

Moreover, we have the following lemma.

Lemma 4.9. *Let $\mathcal{G}_n$ be a potential game with profile space S and P be the transition matrix of the logit dynamics for $\mathcal{G}_n$. For $\beta > 0$, $\emptyset \neq L \subset S$, $\mathbf{x} \in L$ and $0 < \varepsilon < 1$, we have*

$$\mathcal{T}_{S\setminus L}^{\varepsilon}(\mathbf{x}) \leq (B_{\star}^{L})^{-2} \left(\frac{2(1-\varepsilon)}{\varepsilon} + \log \frac{1}{\pi_L(\mathbf{x})}\right),$$

where $\pi_L(\mathbf{x}) = \frac{\pi(\mathbf{x})}{\pi(L)}$ and $B_{\star}^{L} = \min_{\substack{A\subseteq L:\\ \pi(A)\leq 1/2}} B(A)$.

Proof. From Lemma 4.7 we know that the hitting time of $S \setminus L$ can be expressed as a function of the eigenvalues of the matrix $P_{\overline{L}}$. In particular, we have

$$\begin{aligned}
\mathbf{P}_{\mathbf{x}}\left(\tau_{S\setminus L} > t\right) &\leq \exp\left(t \log \lambda_{\max}^{\overline{L}} + \frac{1}{2} \log \frac{1}{\pi_L(\mathbf{x})}\right) \\
\text{(since } 1 - a \leq e^{-a}\text{)} \qquad &\leq \exp\left(-t\left(1 - \lambda_{\max}^{\overline{L}}\right) + \frac{1}{2} \log \frac{1}{\pi_L(\mathbf{x})}\right) \\
\text{(by Lemma 4.5)} \qquad &\leq \exp\left[-\frac{1}{2}\left(t(B_{\star}^{L})^2 - \log \frac{1}{\pi_L(\mathbf{x})}\right)\right] \\
\text{(since } e^{-a} \leq (1+a)^{-1}\text{)} \qquad &\leq \left(1 + \frac{1}{2}\left(t(B_{\star}^{L})^2 - \log \frac{1}{\pi_L(\mathbf{x})}\right)\right)^{-1}.
\end{aligned}$$

Thus, by setting $t = (B_{\star}^{L})^{-2}\left(\frac{2(1-\varepsilon)}{\varepsilon} + \log \frac{1}{\pi_L(\mathbf{x})}\right)$, we have $\mathbf{P}_{\mathbf{x}}\left(\tau_{S\setminus L} > t\right) \leq \varepsilon$ and then $\mathcal{T}_{S\setminus L}^{\varepsilon}(\mathbf{x})$ is upper bounded by this value of t. □

Bounding the hitting time

The following lemma turns out to be useful for proving fast hitting time of profiles not in N.

Lemma 4.10. *Let $\beta \geq \beta_0, \varepsilon > 0$ and let $\mathbf{G}$ be a sequence of n-player potential games asymptotically well-behaved for β. Let p, q the functions generated by $\mathbf{G}$. Then, for each n, at the end of algorithm $\mathcal{A}_{p,q}$ on input $n, \mathcal{G}_n, \beta, \varepsilon$ it holds that for each subset $L \subseteq N$ such that $\pi(L) \leq 1/2$, $B(L) \geq 1/p(n)$.*

Proof. Fix n. It is sufficient to prove that for each R_i chosen by $\mathcal{A}_{p,q}$, its core T_i is non-empty. Indeed, in this case, the algorithm ends only if no subset $L \subseteq N$ such that $\pi(L) \leq 1/2$ has $B(L) > 1/q(n)$. Then, since $\mathbf{G}$ is asymptotically well-behaved, from Definition 4.2 it follows that the last condition is equivalent to $B(L) \geq 1/p(n)$.

As for the non-emptiness of the core, Lemma 4.8 implies that there exists at least one $\mathbf{x} \in R_i$ such that

$$\mathbf{P}_{\mathbf{x}}\left(\tau_{S\setminus R_i} \leq t_{\text{mix}}^{R_i}(\varepsilon)\right) \leq \frac{t_{\text{mix}}^{R_i}(\varepsilon) \cdot B(R_i)}{1 - B(R_i)} \leq \varepsilon,$$

where the last step holds for n sufficiently large since $t_{\text{mix}}^{R_i}$ is at most polynomial by Proposition 4.1 and $B(R_i)$ is at most the inverse of a super-polynomial by hypothesis. □

We are now ready to prove Proposition 4.2

Proof of Proposition 4.2. Fix n. Consider the set of profiles $A_\star \subseteq N$ that minimizes $B(A)$ among all $A \subseteq N$ such that $\pi(A) \leq 1/2$. By Lemma 4.10, $B(A_\star) \geq 1/p(n)$. Moreover, for each n and each $\mathbf{x} \in N$, observe that

$$\log \frac{1}{\pi_N(\mathbf{x})} \leq \log \frac{|S| e^{-\beta \Phi_{\min}}}{e^{-\beta \Phi_{\max}}} \leq \log \frac{e^{n \log m(n)} e^{-\beta \Phi_{\min}}}{e^{-\beta \Phi_{\max}}} = n \log m(n) + \beta\left(\Phi_{\max} - \Phi_{\min}\right),$$

where $\Phi_{\max}$ and $\Phi_{\min}$ denote the maximum and minimum of the potential Φ overall possible strategy profiles. Since $\Phi_{\max} - \Phi_{\min} \leq n \cdot \Delta(n)$ and $\beta \leq \rho(n)/\Delta(n)$, then

$$\log \frac{1}{\pi_N(\mathbf{x})} \leq n \cdot (\log m(n) + \rho(n)) = \rho'(n),$$

where, by assumption on m and ρ, ρ' is a function at most polynomial in its input. Then, for every $\mathbf{x} \in N$, from Lemma 4.9 it follows

$$\mathcal{T}_{S\setminus N}^{\varepsilon}(\mathbf{x}) \leq \left(\frac{1}{B(A_\star)}\right)^2 \cdot \left(\frac{2(1-\varepsilon)}{\varepsilon} + \log \frac{1}{\pi_N(\mathbf{x})}\right) \leq p(n)^2 \cdot \left(\frac{2(1-\varepsilon)}{\varepsilon} + \rho'(n)\right) = \rho_\star(n),$$

where $\rho_\star$ is a function at most polynomial in its input, since p and ρ' are. □

4.3 Application to Specific Games

Let us now introduce some specific classes of games that have been recently studied in the logit dynamics literature. We will show that all these games are asymptotically metastable.

4.3.1 The Curie-Weiss game

Consider the following game-theoretic formulation of the well-studied *Curie-Weiss model* (the *Ising model* on the complete graph), that we will call CW-*game*: each one of n players has two strategies, -1

and $+1$, and the utility of player i for profile $\mathbf{x} = (x_1, \ldots, x_n) \in \{-1, +1\}^n$ is $u_i(\mathbf{x}) = x_i \sum_{j \neq i} x_j$. Observe that for every player i it holds that

$$u_i(\mathbf{x}_{-i}, +1) - u_i(\mathbf{x}_{-i}, -1) = \mathcal{H}(\mathbf{x}_{-i}, -1) - \mathcal{H}(\mathbf{x}_{-i}, +1),$$

where $\mathcal{H}(\mathbf{x}) = -\sum_{j \neq k} x_j x_k$. Hence the CW-game is a potential game with potential function $\mathcal{H}$.

It is known (see, e.g., Chapter 15 in [25]) that the logit dynamics for this game (or equivalently the *Glauber dynamics* for the Curie-Weiss model) has mixing time polynomial in n for $\beta < 1/n$ and super-polynomial as long as $\beta > 1/n$. Moreover, [5] describes metastable distributions for $\beta > c \log n/n$, with c constant, and shows that such distributions are quickly reached from profiles where the number of $+1$ (respectively -1) is a sufficiently large majority, namely if the magnetization k is such that $k^2 > c \log n/\beta$, where the *magnetization* of a profile $\mathbf{x}$ is defined as $M(\mathbf{x}) = \sum_i x_i$.

It has been left open what happens when β lies in the interval $(1/n, c \log n/n)$ in which it is known that the dynamics takes long time to mix, but we do not know if asymptotic metastability occurs, and if a metastable distribution is quickly reached when in the starting point the number of $+1$ is close to the number of -1. We observe that next lemma, along with Theorem 3.1 essentially closes this problem by showing that CW-games are asymptotically metastable for $\beta \geq c/n$ for some constant $c > 1$.

Lemma 4.11. *Let $\mathbf{G}$ be a sequence of n-player CW-games. Then the logit dynamics for $\mathbf{G}$ is asymptotically metastable for every $\beta > c/n$, for constant $c > 1$.*

Proof. We will next prove that $\mathbf{G}$ is partitioned by the logit dynamics for every $\beta > c/n$. The claim then follows from Theorem 3.1.

Fix n and let S_+ (resp., S_-) be the set of profiles with positive (resp., negative) magnetization in $\mathcal{G}_n$. Let us set $R_1 = S_+$ and $R_2 = S_-$. It is known that the bottleneck ratio of these subset is super-polynomial for every $\beta > c/n$, for constant $c > 1$, (see, e.g., Chapter 15 in [25]). Moreover, in [24] it has been proved that the mixing time of the chain restricted to S_+ (resp. S_-) is actually $c_1 n \log n$ for some constant $c_1 > 0$[10].

Let now ζ be the unique positive root [14, 15] of the function

$$f(x) = \frac{e^{\beta x}(1-x) - e^{-\beta x}(1+x)}{e^{\beta x}(1-x) + e^{-\beta x}(1+x)}.$$

Observe that $\zeta \in [0, 1]$, it is non decreasing in β and does not depend on n. Let now Z_+ be the set of profiles with magnetization $k \geq \zeta n$ and Z_- be the set of profiles with magnetization $k \leq -\zeta n$. Note that for a constant $c > 1$ and n sufficiently large, we have that $|k| \geq 1$ [15].

Consider now the following partition of S: $T_1 = Z_+$, $T_2 = Z_-$ and $N = S \setminus (Z_+ \cup Z_-)$. We prove that from every profile $\mathbf{x} \in Z_+$ the dynamics hits a profile $\mathbf{y} \in S_-$ in a time equivalent to the mixing time of the chain restricted to S_+ with probability at most ε. Consider, indeed, the magnetization chain, i.e., the birth and death chain on the space $\{-n, 2-n, \ldots, n-2, n\}$. Then we are interested in the hitting time τ_l of $l \leq 0$ when the starting point is k. Clearly, in order to reach magnetization l it is necessary to reach magnetization k', with $l < k' < k$. And for reaching l from k' it is necessary to reach k'' such that $l \leq k'' < k'$. Then, we show that there is k' from which the chain quickly goes back to $n\zeta$ with high probability without ever hitting the profile k''. In particular, in [14, Theorem 4.10] it has been showed that there are k' and k'' such that

$$\mathbf{P}_{k'}\left(\tau_{n\zeta} \leq c_2 n \log n\right) \geq 1 - o(1) \qquad \text{and} \qquad \mathbf{P}_{k'}\left(\tau_{k''} \geq c_2 n \log n\right) \geq 1 - o(1).$$

[10]The result in [24] refers to censored chains, that are exactly the same as our restricted chain, except that the probability that the original chain from a profile $\mathbf{x}$ goes out from L is "reflected" to some profile in L different from $\mathbf{x}$, instead than being "added" to the probability to stay in $\mathbf{x}$. It is immediate to see how their result extends also to our restricted chains.

Hence, it follows that

$$\mathbf{P}_k\left(\tau_l \geq c_1 n \log n\right) \geq (1 - o(1))^{\kappa} = 1 - o(1),$$

where κ is a constant depending only on c_1 and c_2. (Clearly, everything holds symmetrically by considering the set Z_-.)

Finally, observe that from [14] (Theorems 4.4, 4.9 and 4.10 – see also [15]), we have that for each profile $\mathbf{x} \in N$, the hitting time of $T_1 \cup T_2$ is polynomial with high probability. □

The asymptotic metastability of the Curie-Weiss game was studied in [5], together with pure coordination games, and graphical coordination games on the ring. It is not hard to see that previously known results about these last two games can be easily derived within our framework.

4.3.2 A Congestion Game

We now exemplify how our framework can be used to prove metastability for another class of games, by sketching a proof that the following game inspired by the well-known Pigou's congestion game is asymptotically metastable.

There are two links, one has fixed cost $z = k - \varepsilon$ with $k \in \{1, \ldots, n-1\}$ and $\varepsilon > 0$, whereas the second one has congestion-dependent cost $c(\mathbf{x})$, where $c(\mathbf{x})$ denotes the number of players choosing this second linkin the profile $\mathbf{x}$. It is well known that this game is a potential game [34], with potential function $\Phi(\mathbf{x}) = \left[z(n - c(\mathbf{x})) + \sum_{i=1}^{c(\mathbf{x})} i\right] = \frac{1}{2}\left(2zn - 2zc(\mathbf{x}) + c(\mathbf{x})(c(\mathbf{x}) + 1)\right)$. We next show that the logit dynamics for this game is partitioned, from which it follows that the game is asymptotically metastable.

Consider, indeed, all profiles such that there are k players on the second link, i.e., the one with congestion dependent cost. Observe that there are $\binom{n}{k} > 1$ of these profiles, and for each of them we define R_i that contains only this profile. Observe that the inverse of bottleneck ratio of this profile is $e^{-\beta\varepsilon}$ that is super-polynomial for β sufficiently large (it is easy to see that for smaller β the mixing time is polynomial). Since these sets are singletons, it is immediate that their mixing time is at most polynomial and it must hard to leave this set from $T_i = R_i$. It is only left to prove that the logit dynamics converges to one of these sets quickly from every other profile. However, this immediately follows by a simple birth-and-chain argument, by observing that, for large β, the number of players adopting the congestion-dependent link increases with large probability if $c(\mathbf{x}) < k$ and decrease with high probability if $c(\mathbf{x}) > k$.

4.3.3 Opinion Formation Games on Social Networks

Another class of games that has been recently analyzed under the lens of the logit dynamics has been the class of opinion formation games [18]. Here, agent i must express an opinion x_i from a discrete set. Let us assume for simplicity that the available opinions are only $\{0, 1\}$. Agent i is moreover provided with a private belief $b_i \in [0, 1]$. Agents are located on vertices of a graph representing a social network. Then, agent's opinion is the compromise between her own private belief on the topic at the hand, and the public opinion expressed by "friends" on the social network. This compromise is represented through the utility function of agents, that involve two components: the first one, $(x_i - b_i)^2$, measures the distances between agent i's public opinion and her own private belief, whereas the second component, $\sum_{j \in N(i)} (x_i - x_j)^2$, measures the distance between agent i's public opinion and the opinion of each agent j that is a neighbor of i in the social network. Opportune weights may be set to balance these two components and also to balance the contributions of different neighbors.

To exemplify the application of our framework to these games, let us consider a specific setting, in which $b_i = 1/2$ and the underlying network is $K_{m,m}$, i.e., the complete bipartite graph with m nodes in each side. We next observe that this game is partitioned by the logit dynamics, and thus it enjoys

asymptotic metastability. To this aim, let us define R_0 as the subset of profiles $\mathbf{x}$ such that there is a path $\mathbf{x}_0, \ldots, \mathbf{x}_k$, with $\mathbf{x}_0 = (0, \ldots, 0)$, $\mathbf{x}_k = \mathbf{x}$, $\mathbf{x}_\ell$ and $\mathbf{x}_{\ell-1}$ diverging only in the opinion of a single agent, and in each profile of the path there are less than κ edges whose endpoint have discording edges, where $\kappa = \lceil m^2/2 \rceil$. Similarly one can define R_1. It follows from [18] that these two sets have super-polynomial bottleneck ratio for β sufficiently large. Moreover, it is not hard to see that the mixing time of the chain restricted to these sets is polynomial. This can be proved by adapting an argument used in [3] for bounding the mixing time of games with dominant strategies. Finally, with arguments similar to the one described above in Lemma 4.11, one can see that there are T_0 and T_1 from which it is hard to leave R_0 and R_1, respectively, but that are easy to reach from each remaining profile.

5 Additional applications

While the main objective of this work is to prove that for every potential game the behavior of the logit dynamics is always asymptotically metastable, we observe that metastability and the tool that we proposed in this work, namely the concept of partitioned Markov chain, may be of independent interest and of broader application also outside game theory. Below we sketch a setting in which the corresponding Markov chain appears to be partitioned, and one in which the resulting metastable behavior may play a fundamental role.

5.1 Graph Clustering Algorithms

Given a graph $G = (V, E)$ with maximum degree bounded by d, and a subset $S \subseteq V$, the *outer conductance* of S is the ratio between the number of edges between S and $V \setminus S$, and $d|S|$. The *inner conductance* of S is the minimum among all subset $T \subseteq S$ with $|T| \leq |S|/2$ of $\frac{\#\text{edges}(T, S \setminus T)}{d|T|}$. Given a parameter ε, the graph G is defined to be (k, ϕ)-*clusterable*, if it can be partitioned into no more than k parts, such that the inner conductance of the induced subgraph on each part is at least ϕ and the outer conductance of each part is at most $c_{d,k}\varepsilon^4\phi^2$, where $c_{d,k}$ depends only on d and k.

Czumaj et al. [12] studied the problem of recognizing the cluster structure of a graph in the framework of property testing. Specifically, they present a sublinear algorithm that takes in input a graph with bounded degree d and parameters k, ϕ, and ε and tests if the graph is either (k, ϕ)-clusterable or we need to add or delete more than εdn edges to obtain a (k, ϕ^*)-clusterable graph, where $\phi^* = c'_{d,k}\frac{\phi^2\varepsilon^4}{\log n}$ and $c'_{d,k}$ depends only on d and k.

The algorithm proposed is roughly as follows: run a random walk on the graph for about $\frac{\log n}{\phi^2}$ steps starting from $k \log k$ starting nodes chosen uniformly at random; if two different walks ends up with a close distribution, then the algorithm assigns them to the same cluster, otherwise they are assigned to different clusters. In other words, the algorithm is exploiting the asymptotic metastability of the random walk on a graph with bounded degree d with respect to a class C of functions that are logarithmic in n and quadratic in $1/\phi$: for every starting point of the walk, it mixes within this time frame in a subgraph whose inner conductance is at least ϕ and whose outer conductance is at most $c_{d,k}\varepsilon^4\phi^2$, and remains therein for an amount of time that is at least the square of this time frame (when the dynamics starts outside any clusters, or at very border of it, the convergence is not guaranteed to a single cluster but to a combination of them; still the algorithm of [12] deals this case by discarding these starting points).

In particular, this metastable behavior follows from our Theorem 3.1 since the random walk is partitioned with respect to these functions: if there are k clusters with outer conductance $c_{d,k}\varepsilon^4\phi^2$, then there are $R_1, \ldots, R_k$ with bottleneck ratio at least $1/\phi^4$; Moreover, if these clusters have inner conductance at most ϕ, then the corresponding restricted chain mixes in time $O\left(\frac{\log n}{\phi}\right)$ and there is a non-empty subset of profiles from which it is hard to leave the subset within this time frame.

Clearly, most of the work in [12] is in how to exploit this metastable behavior to design an algorithm with the desired properties. We believe, however, that Theorem 3.1 can be an useful tool for the analysis of randomized algorithms and, it may suggest that the idea developed in a setting (e.g., the idea of [12] about graph clustering) may be useful to other settings in which a similar metastable behavior holds. Such an example can be found in the very recent work of Cruciani et al. [11]. They have analyzed the 2-choice dynamics, according to which each node can be in one of two states, and at each time step each node u randomly selects a pair of neighbors, and, if these neighbors have the same state, then u copies it. They proved that on graphs consisting of two well-separated clusters, the 2-choice dynamics enjoys asymptotic metastability. They in fact prove that the corresponding Markov chain is partitioned (even if they do not explicitly use this terminology): there are four classes of configurations, the ones in which almost every node in both the clusters of the graph have the same state, and the ones in which nodes in different clusters have different states; they then focus only on the latter ones, and they show that the dynamics remains within these configurations for polynomial time, whereas these configurations are quickly reached (in logarithmic time) as long as the starting point of the chain is sufficiently far away from the configurations in which both clusters have the same state.

5.2 Random Network Generation

Several varieties of random graph models have been developed in recent years to explain characteristic of observed real-world networks. A key model, extensively used in sociology literature [30, 37], is the *exponential random graph* model. The model is defined in terms of the number of subgraphs X (e.g., number of edges, or number of triangles) contained in the network; each of these subgraph has a weight β_i and the number of occurrence of each subgraph and their weights define the so-called *graph Hamiltonian* $H(G)$; the model then returns a graph G with probability $P(G) \approx e^{H(G)}$.

Sampling from this distribution is crucial for parameter estimation property testing, and for adopting this model in experimental analysis of networks. Typically, sampling has been carried out using the Glauber dynamics procedure. In the implementation of this procedure, a very important role is played by weights β_i. In particular, Bhamidi et al. [6] distinguish the behavior of the Glauber dynamics in two regimes: the high temperature regime in which essentially there is a unique local minimum in the graph Hamiltonian, and the low temperature regime in which the Hamiltonian has more than one local minimum.

Despite the relevance of this network model and the widespread use of Glauber dynamics to sample from this model, this practice has been strongly criticized in [6]. Indeed, therein it has been observed that the Glauber dynamics quickly converges to the desired distribution in the high temperature regime, but in this case the resulting distribution is not significantly different from the one resulting from applying the simpler, and faster, Erdos-Renyi graph model, known to violate many desired properties.

Moreover, Bhamidi et al. [6] also show that in the low temperature regime the Glauber dynamics takes exponential time to converge to the desired distribution, and hence conclude that this procedure is infeasible in this regime. In particular, their argument shows that when the Glauber dynamics starts close to one of the local minima of the graph Hamiltonian, then it takes exponential time to leave this neighborhood. In other words, there are subsets of states of the chain with bottleneck ratio at most the inverse of an exponential function, and there are states within these subsets from which it takes exponential time to leave the subset. These properties suggest that the dynamics may be partitioned and thus, by Theorem 3.1, enjoy asymptotic metastability.

This sheds new light on the convergence of the Glauber dynamics to the desired distribution: even if the convergence to the stationary distribution may take long time starting from specific states, by proving the dynamics is partitioned, one would know that a (meta)stable distribution is still reached from every starting point. Then one may ask how "good" is this distribution (e.g., how far is it from the Erdos-Renyi model) or if there are classes of starting configurations from which the dynamics reaches

good metastable distributions. In other words, our framework may open new avenues in the analysis of sampling techniques, showing they partially work even when their mixing time is large.

6 Conclusions and open problems

In this work we prove that for every asymptotically well-behaved potential game and every starting point of this game there is a distribution that is metastable for super-polynomial time and it is quickly reached. Even if our definition of asymptotically well-behaved potential games was introduced to model the games for which it makes sense to give asymptotic results about the logit dynamics, one may wonder if our techniques captures all the cases in which asymptotically metastability is possible. For example, we know that the larger counterpart of asymptotically-well behaved games defined in Appendix C is sufficient, but is it also *necessary*? The main obstacle for proving this direction consists in the absence of any tool for proving or disproving metastability of distributions that are largely different from the ones considered in this work. Also, given that our arguments are game-independent, it would be interesting to see whether sufficient and necessary conditions can be refined for specific subclasses of games.

Our convergence rate results hold if β is small enough. As we mention above, an assumption on β is in general necessary because when β is high enough logit dynamics roughly behaves as best-response dynamics; moreover, the convergence in polynomial-time of best-response dynamics for potential games is known to be hard [17]. Interestingly, this difference in the behavior of the logit dynamics for different values of β suggests that "the more noisy the system is, the more (meta)stable it is."

Our result is in a sense existential, since it is not practical to explicitly describe the distributions via the execution of Algorithm 4.1. It is then an interesting open problem to characterize the sets R_i's and T_i's returned by this algorithm for some specific classes of games in order to understand better the stability guarantee of the distributions. A better understanding of spectra of the transition matrix along the lines of the results we prove in Appendix D may help in answering some of the questions above.

Naturally, there are other questions of general interest about metastability that we do not consider. For example, akin to price of anarchy and price of stability, one may ask what is the performance of a system in a metastable distribution? One might also want to investigate metastable behavior of different dynamics in potential games, such as best-response dynamics. However, in the latter case, no matter what selection rule is used to choose which player has to move next, a profile is never visited twice in time since at each step the potential goes down. Therefore, the "transient" behavior of best-response dynamics would roughly correspond to a (possibly exponentially long) sequence of profiles visited. This, however, would not add much to our understanding of the transient phase of best-response dynamics.

Finally, we believe that the concept of asymptotic metastability and tool of partitioned chain that we introduced in this work, may be of independent interest and with broad applications in algorithm design and analysis: we highlighted a few of these applications in Section 5, but we believe that many other applications are just behind the corner.

Acknowledgments

We wish to thank Paul W. Goldberg for many invaluable discussions related to a number of results discussed in this paper, and an anonymous reviewer for the enlightening comments on an earlier version of this work. Diodato Ferraioli was supported by the "GNCS – INdAM". Carmine Ventre was supported by EPSRC, through grant EP/M018113/1.

References

[1] C. Alós-Ferrer and N. Netzer. The logit-response dynamics. *Games and Economic Behavior*, 68(2):413 – 427, 2010.

[2] A. Asadpour and A. Saberi. On the inefficiency ratio of stable equilibria in congestion games. In *Proc. of the 5th International Workshop on Internet and Network Economics (WINE'09)*, volume 5929 of *Lecture Notes in Computer Science*, pages 545–552. Springer, 2009.

[3] V. Auletta, D. Ferraioli, F. Pasquale, P. Penna, and G. Persiano. Convergence to equilibrium of logit dynamics for strategic games. *Algorithmica*, 76(1):110–142, 2016.

[4] V. Auletta, D. Ferraioli, F. Pasquale, and G. Persiano. Mixing time and stationary expected social welfare of logit dynamics. *Theory of Computing Systems*, 53(1):3–40, 2013.

[5] V. Auletta, D. Ferraioli, F. Pasquale, and G. Persiano. Metastability of logit dynamics for coordination games. *Algorithmica*, Sep 2017.

[6] S. Bhamidi, G. Bresler, and A. Sly. Mixing time of exponential random graphs. In *49th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2008, October 25-28, 2008, Philadelphia, PA, USA*, pages 803–812, 2008.

[7] L. E. Blume. The statistical mechanics of strategic interaction. *Games and Economic Behavior*, 5:387–424, 1993.

[8] A. Bovier. Metastability: a potential theoretic approach. In *Proc. of the International Congress of Mathematicians*, volume III, pages 499–518. European Mathematical Society, 2006.

[9] X. Chen, X. Deng, and S. Teng. Computing nash equilibria: Approximation and smoothed complexity. In *FOCS*, pages 603–612, 2006.

[10] S. Chien and A. Sinclair. Convergence to approximate nash equilibria in congestion games. In *SODA*, pages 169–178, 2007.

[11] E. Cruciani, E. Natale, and G. Scornavacca. On the Metastability of Quadratic Majority Dynamics on Clustered Graphs and its Biological Implications. *ArXiv e-prints*, 2018.

[12] A. Czumaj, P. Peng, and C. Sohler. Testing cluster structure of graphs. In *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing, STOC 2015, Portland, OR, USA, June 14-17, 2015*, pages 723–732, 2015.

[13] C. Daskalakis. On the complexity of approximating a nash equilibrium. In *SODA*, pages 1498–1517, 2011.

[14] J. Ding, E. Lubetzky, and Y. Peres. Censored glauber dynamics for the mean field ising model. *Journal of Statistical Physics*, 137(3):407–458, 2009.

[15] J. Ding, E. Lubetzky, and Y. Peres. The mixing time evolution of glauber dynamics for the mean-field ising model. *Communications in Mathematical Physics*, 289(2):725–764, 2009.

[16] G. Ellison. Learning, local interaction, and coordination. *Econometrica*, 61(5):1047–1071, 1993.

[17] A. Fabrikant, C. H. Papadimitriou, and K. Talwar. The complexity of pure nash equilibria. In *STOC*, pages 604–612, 2004.

[18] D. Ferraioli, P. W Goldberg, and C. Ventre. Decentralized dynamics for finite opinion games. *Theoretical Computer Science*, 648:96–115, 2016.

[19] D. Ferraioli and C. Ventre. Metastability of asymptotically well-behaved potential games - (extended abstract). In *Mathematical Foundations of Computer Science 2015 - 40th International Symposium, MFCS 2015, Milan, Italy, August 24-28, 2015, Proceedings, Part II*, volume 9235 of *Lecture Notes in Computer Science*, pages 311–323, 2015.

[20] F. Hollander. Metastability under stochastic dynamics. *Stochastic Processes and their Applications*, 114(1):1–26, 2004.

[21] F. Hollander. Three lectures on metastability under stochastic dynamics. In *Methods of Contemporary Mathematical Statistical Physics*, volume 1970 of *Lecture Notes in Mathematics*, pages 1–24. Springer Berlin / Heidelberg, 2009.

[22] R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, 1990.

[23] M. J. Kearns and Y. Mansour. Efficient nash computation in large population games with bounded influence. In *UAI*, pages 259–266, 2002.

[24] D. Levin, M. Luczak, and Y. Peres. Glauber dynamics for the mean-field ising model: cut-off, critical power law, and metastability. *Probability Theory and Related Fields*, 146:223–265, 2010.

[25] D. Levin, Y. Peres, and E. L. Wilmer. *Markov Chains and Mixing Times*. American Mathematical Society, 2008.

[26] R. J. Lipton, E. Markakis, and A. Mehta. Playing large games using simple strategies. In *ACM Conference on Electronic Commerce*, pages 36–41, 2003.

[27] R. D. McKelvey and T. R. Palfrey. Quantal response equilibria for normal form games. *Games and economic behavior*, 10(1):6–38, 1995.

[28] D. Monderer and L. S. Shapley. Potential games. *Games and Economic Behavior*, 14(1):124 – 143, 1996.

[29] A. Montanari and A. Saberi. Convergence to equilibrium in local interaction games. In *Proc. of the 50th Annual Symposium on Foundations of Computer Science (FOCS'09)*. IEEE, 2009.

[30] M. Newman. The structure and function of complex networks. *SIAM review*, 45(2):167–256, 2003.

[31] E. Olivieri and M. E. Vares. *Large deviation and metastability*. Cambridge University Press, 2005.

[32] S. Osnaga. On rank one matrices and invariant subspaces. *Balkan Journal of Geometry and Its Applications*, 10(1):145, 2005.

[33] H. Peyton Young. *The diffusion of innovations in social networks*, chapter in "The Economy as a Complex Evolving System", vol. III, Lawrence E. Blume and Steven N. Durlauf, eds. Oxford University Press, 2003.

[34] R. W. Rosenthal. A class of games possessing pure-strategy nash equilibria. *International Journal of Game Theory*, 2(1):65–67, 1973.

[35] R. J. Schiller. *Irrational Exuberance*. Wiley, 2000.

[36] A. Skopalik and B. Vöcking. Inapproximability of pure nash equilibria. In *STOC*, pages 355–364, 2008.

[37] T. Snijders, P. Pattison, G. Robins, and M. Handcock. New specifications for exponential random graph models. *Sociological methodology*, 36(1):99–153, 2006.

A Background material

In this section, we give some background on a couple of techniques adopted to prove our results: coupling of Markov chains, and tools to bound the hitting time.

A.1 Markov chain coupling

A *coupling* of two probability distributions μ and ν on a state space S is a pair of random variables (X, Y) defined on $S \times S$ such that the marginal distribution of X is μ and the marginal distribution of Y is ν. A *coupling of a Markov chain* $\mathcal{M}$ on S with transition matrix P is a process $(X_t, Y_t)_{t=0}^{\infty}$ with the property that X_t and Y_t are both Markov chains with transition matrix P. Similarly, a *coupling of Markov chains* $\mathcal{M}$, $\bar{\mathcal{M}}$ both defined on S with transition matrices P and $\bar{P}$, respectively, is a process $(X_t, Y_t)_{t=0}^{\infty}$ with the property that X_t is a Markov chain with transition matrix P and Y_t is a Markov chain with transition matrix $\bar{P}$.

When the two coupled chains start at $(X_0, Y_0) = (\mathbf{x}, \mathbf{y})$, we write $\mathbf{P}_{\mathbf{x},\mathbf{y}}(\cdot)$ for the probability of an event on the space $S \times S$. The following theorem establishes the importance and the usefulness of the coupling.

Theorem A.1 (Coupling (see, e.g., Proposition 4.7 and Theorem 5.2 in [25]))**.** *Let $\mathcal{M}$, $\bar{\mathcal{M}}$ be two Markov chains with finite state space S and transition matrices P and $\bar{P}$, respectively. For each pair of states $\mathbf{x}, \mathbf{y} \in S$ consider a coupling (X_t, Y_t) of $\mathcal{M}$ and $\bar{\mathcal{M}}$ with starting states $X_0 = \mathbf{x}$ and $Y_0 = \mathbf{y}$. Then*

$$\left\| P^t(\mathbf{x}, \cdot) - \bar{P}^t(\mathbf{y}, \cdot) \right\|_{\mathrm{TV}} \leq \mathbf{P}_{\mathbf{x},\mathbf{y}}\left(X_t \neq Y_t\right).$$

This theorem is mainly used for bounding the distance of a Markov Chain to the stationary distribution.

A.2 Hitting time tools

Consider a reversible Markov chain with state space S and transition matrix P. For $L \subseteq S$ let $P_{\overline{L}}$, $\lambda_i^{\overline{L}}$ and $\lambda_{\max}^{\overline{L}}$ as defined in Section 4.2.3. Here we give a well known (see, e.g., [29]) variational characterization of $\lambda_{\max}^{\overline{L}}$ as expressed by the following lemma.

Lemma A.1. *Consider a reversible Markov chain with state space S, transition matrix P and stationary distribution π. For every $L \subseteq S$ we have*

$$1 - \lambda_{\max}^{\overline{L}} = \inf_{\varphi} \frac{\mathcal{E}_P(\varphi)}{\mathbf{E}_\pi\left[\varphi^2\right]},$$

where $\mathcal{E}_P(\varphi)$ is defined as in (8)*, $\mathbf{E}_\pi\left[\varphi^2\right] = \sum_{\mathbf{x}} \pi(\mathbf{x})\varphi^2(\mathbf{x})$ and the inf is taken over functions φ such that $\varphi(\mathbf{x}) = 0$ for $\mathbf{x} \in S \setminus L$ and $\mathbf{E}_\pi\left[\varphi^2\right] \neq 0$.*

Since the statement of Lemma 4.7 is slightly different from the ones found in previous literature, we attach a proof for sake of completeness.

Proof of Lemma 4.7. Let φ_L be the characteristic function on L, that is $\varphi_L(\mathbf{x}) = 1$ if $\mathbf{x} \in L$ and 0 otherwise. Then

$$\mathbf{P}_{\mathbf{x}}\left(\tau_{S\setminus L} > t\right) = \sum_{\mathbf{y}\in S} P_{\overline{L}}^t(\mathbf{x}, \mathbf{y}) = \sum_{\mathbf{y}\in S} P_{\overline{L}}^t(\mathbf{x}, \mathbf{y})\varphi_L(\mathbf{y}) = (P_{\overline{L}}^t\varphi_L)(\mathbf{x}). \tag{9}$$

Since $P_{\overline{L}}$ is reversible with respect to π_L, we have that its eigenvectors, $\psi_1, \ldots, \psi_{|S|}$, form an orthonormal basis with respect to the inner product $\langle \cdot, \cdot \rangle_{\pi_L}$: in particular we can write $\varphi_L = \sum_i \alpha_i \psi_i$, where $\sum_i \alpha_i = 1$ and each $\alpha_i \geq 0$. Hence and from the linearity of the inner product we have

$$\begin{aligned} \langle P_{\overline{L}}^t \varphi_L, P_{\overline{L}}^t \varphi_L \rangle_{\pi_L} &= \sum_i \sum_j \langle \alpha_i \left(\lambda_i^{\overline{L}}\right)^t \psi_i, \alpha_j \left(\lambda_j^{\overline{L}}\right)^t \psi_j \rangle_{\pi_L} \\ \text{(by orthogonality)} &= \sum_i \left(\lambda_i^{\overline{L}}\right)^{2t} \langle \alpha_i \psi_i, \alpha_i \psi_i \rangle_{\pi_L} \\ &\leq \left(\lambda_{\max}^{\overline{L}}\right)^{2t} \langle \varphi_L, \varphi_L \rangle_{\pi_L} = \left(\lambda_{\max}^{\overline{L}}\right)^{2t}, \end{aligned} \tag{10}$$

where the last equality follow from the definition of φ_L. Moreover,

$$\pi_L(\mathbf{x})[(P_{\overline{L}}^t \varphi_L)(\mathbf{x})]^2 \leq \sum_{\mathbf{y} \in S} \pi_L(\mathbf{y})[(P_{\overline{L}}^t \varphi_L)(\mathbf{y})]^2 = \langle P_{\overline{L}}^t \varphi_L, P_{\overline{L}}^t \varphi_L \rangle_{\pi_L}. \tag{11}$$

The theorem follows from (9), (10), (11). □

B Asymptotically Well-Behaved Games: The Route to the Definition

In this section we will incrementally identify all the necessary technical properties that one needs to model sequences of games for which the behavior of the logit dynamics is asymptotically well-defined. We will show how this leads to our Definition 4.2.

B.1 First Tentative

Let us start by assuming that all potential games are asymptotically well-behaved. Lemma B.1 shows that this is not the case and we need a more restrictive definition.

Lemma B.1. *There is a sequence of n-player potential games $\mathbf{G}$ which is not asymptotically metastable for every β sufficiently high and every $\varepsilon < \frac{1}{4}$.*

Proof. We will show a sequence of n-player games $\mathbf{G}$ such that for every $0 < \varepsilon < 1/4$, for infinitely many value of n and for each polynomial p in n and each super-polynomial q in n, there is a profile $\mathbf{x}$ of $\mathcal{G}_n$ such that the logit dynamics for $\mathcal{G}_n$ does not converge in time at most $p(n)$ from $\mathbf{x}$ to any $(\varepsilon, q(n))$-metastable distribution, even if the mixing time of the dynamics is larger than $p(n)$.

Consider the following pairs (p_j, q_j), where $p_j = n^j$ and $q_j = \exp\left(\log n \cdot \log^{(j)} n\right)$, where $\log^{(j)}$ is the j-th functional iteration of the logarithm function. Let us denote as n_j a value such that $p_j(n_j) < q_j(n_j) - \varepsilon$. Such a value surely exists since p is polynomial and q is super-polynomial. Moreover, observe that for each $n > n_j$, we have $p_j(n) < q_j(n) - \varepsilon$. Thus, we can assume without loss of generality that $1 = n_0 < n_1 < n_2 < \ldots$. Now let $\mathcal{T}$ be a function that is asymptotically sandwiched between p_j and q_j, for every j. This can be guaranteed by letting $\mathcal{T}$ be a function such that $\mathcal{T}(n) = q_j(n) - \varepsilon$ for j such that $n_j < n \leq n_{j+1}$. Note that for every p_j and for every $n \geq n_j$, we have $\mathcal{T}(n) = q_k(n) - \varepsilon > p_k(n) \geq p_j(n)$, where $k \geq j$ is such that $n_k \leq n < n_{k+1}$. Similarly, for every q_j and for every $n \geq n_j$ we have $\mathcal{T}(n) = q_k(n) - \varepsilon < q_k(n) \leq q_j(n)$. The situation is depicted in Figure 1.

Let now $\mathbf{G}$ be a sequence of n-player potential games such that, for each $\mathcal{G}_n$, each player has exactly two strategies, say 0 and 1. Consider the potential function Φ of $\mathcal{G}_n$ such that for every $t = 0, \ldots, n-1$ and every profile $\mathbf{x}$ wherein exactly t players play strategy 1 we have $\Phi(\mathbf{x}) = n - t$, while $\Phi(1, \ldots, 1) = 1 + k_n$, where $k_n = \frac{1}{\beta} \log\left(\frac{\mathcal{T}(n)}{\varepsilon} - 1\right)$, β being the rationality parameter of the logit dynamics.

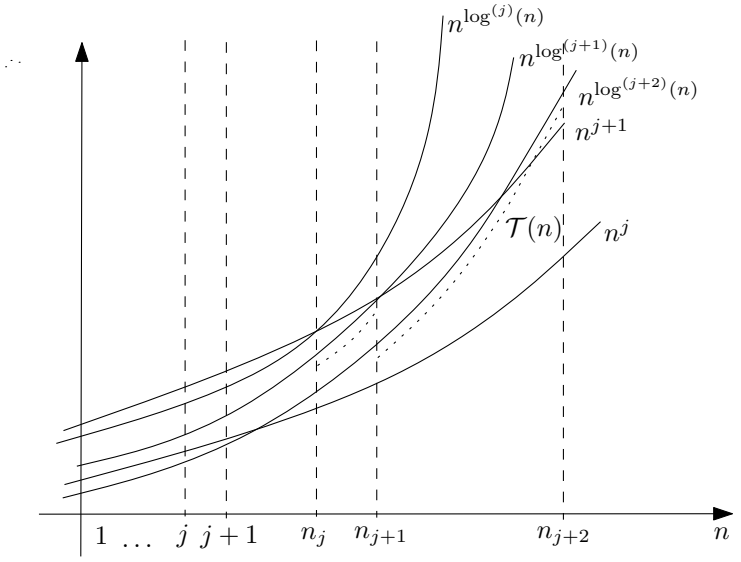

Figure 1: The figure shows how $\mathcal{T}(n)$ is built around the functions p_j's and q_j's so that it is hard to classify $\mathcal{T}(n)$ either as a polynomial or as a super-polynomial. Note that $\mathcal{T}(n)$ is the inverse of the bottleneck ratio of profile $(1,\dots,1)$ and thus it describes the time needed to leave that profile.

Observe that if there is a pair (p,q) with p polynomial in n and q super-polynomial in n such that it is possible to prove that the logit dynamics for $\mathcal{G}_n$ is asymptotic metastable with parameters $p(n)$ and $q(n)$, then there is $j^\star$ such that the results holds also with $(p_{j^\star}(n), q_{j^\star}(n))$ in place of $p(n)$ and $q(n)$, where $(p_{j^\star}, q_{j^\star})$ corresponds to one of the pair of functions described above. Hence, in order to prove the lemma is sufficient to show that it holds only for pairs $(p_j(n), q_j(n))$ as described above.

Note that, by taking β sufficiently high, we have that: (i) $\pi(0,\dots,0) \geq \frac{1}{2}$; (ii) there exists a j such that, for every subset $L \subseteq \{0,1\}^n \setminus \{(0,\dots,0),(1,\dots,1)\}$, the bottleneck ratio $B(L)$ is at least the inverse of p_j; (iii) the bottleneck ratio $B(1,\dots,1) = \frac{\varepsilon}{\mathcal{T}(n)}$.

Firstly note that the mixing time of logit dynamics for $\mathcal{G}_n$ is not polynomial. Indeed, from Theorem 2.2, it follows that the mixing time is at least $\frac{\mathcal{T}}{4\varepsilon}$. However, as suggested above, for each p_j, we have $\frac{\mathcal{T}}{4\varepsilon} > \mathcal{T} > p_j$ for infinitely many n, and hence the mixing time is asymptotically greater than every polynomial p_j.

We next discuss that no metastable distribution is stable for a super-polynomial time or, even if there is one, it cannot be reached in polynomial time. From Lemma 2.1 and Lemma 2.2, we have that for each n the distribution π_1 that assigns probability 1 to the profile $(1,\dots,1)$ is $(\varepsilon, \mathcal{T}(n))$-metastable. However, as suggested above, for each q_j, the function $\mathcal{T}$ is smaller than q_j for infinitely many n. Thus, the distribution π_1 is metastable for time that is asymptotically smaller than every super-polynomial q_j. Note that this argument extends to every $(3\varepsilon, \mathcal{T}(n))$-metastable distribution μ that is within distance 2ε from π_1. Finally, observe that, the remaining distributions that are far from π_1 cannot be reached quickly from $(1,\dots,1)$. In fact, from Lemma 4.8 below, for every polynomial p_j the probability that the logit dynamics leaves the profile $(1,\dots,1)$ in p_j steps, is at most $\frac{\varepsilon \cdot p_j}{\mathcal{T}-\varepsilon} < \varepsilon$, for n sufficiently large. Hence, for every p_j, starting from $(1,\dots,1)$ the pseudo-mixing time of every distribution μ that is at least 2ε-far from π_1 is asymptotically greater than p_j. □

Remark B.1. The game described in the proof of Lemma B.1 also shows the necessity of having a definition of asymptotic metastability as the one given in Definition 2.3.

Consider, indeed, the weaker definition of asymptotic metastability in which for each n there is a polynomial $p_n(n)$ and a super-polynomial $q_n(n)$ governing convergence and stability time of metastability, respectively (i.e., a definition in which the order of quantifiers is reversed). This concept might

at first glance look meaningful. However, it is instead of scarce significance as the distinction between polynomials and super-polynomials might become null in the limit.

The game in Lemma B.1 exemplifies this phenomenon, since it does not satisfies the metastability notion given in the one in Definition 2.3, but it satisfies this weak notion. Indeed, for every n, there is a super-polynomial function, namely $\tilde{q}_n = q_j - \varepsilon$ for j such that $n_{j-1} < n \leq n_j$, such that π_1 is $(\varepsilon, \tilde{q}_n(n))$-metastable. Obviously, the pseudo-mixing time of this distribution from the profile $(1, \ldots, 1)$ is 1. From the remaining profiles, the dynamics quickly converges to the stationary distribution for every β sufficiently large (this follows from well-known results about birth-and-death chains).

B.2 Second Tentative

The game described in the proof of Lemma B.1 is clearly a game for which it makes no sense to prove asymptotic results about the behavior of the logit dynamics, since this object changes infinitely often. Hence, we need to constrain the set of games of interest in this work.

To get an intuition of the condition that we are going to define, it is worth to look more closely at the game of Lemma B.1. This game necessitates the update of the function $\mathcal{T}$ infinitely often when adding a new player. This update is done so that the new profile $(1, \ldots, 1)$ has a bottleneck ratio that cannot be described by any of the functions considered at that point. This process is never ending and gives no asymptotic meaning to $\mathcal{T}$ in the limit.

The intuition is then that it does not make sense to prove asymptotic results about metastability of the logit dynamics, whenever we need infinitely many functions for describing the bottleneck ratio of a single profile as n changes, so that we are never able to classify the bottleneck ratio either as polynomial or as super-polynomial. Recall that, for every profile $\mathbf{x}$, its bottleneck is defined as follows:

$$\begin{aligned} B(\mathbf{x}) &= \sum_i \sum_{s_i \in S_i} P(\mathbf{x}, (\mathbf{x}_{-i}, s_i)) = \frac{1}{n} \sum_i \sum_{\substack{s_i \in S_i \\ s_i \neq x_i}} \frac{e^{-\beta \Phi(\mathbf{x}, (\mathbf{x}_{-i}, s_i))}}{\sum_{z_i \in S_i} e^{-\beta \Phi(\mathbf{x}_{-i}, z_i)}} \\ &= 1 - \frac{1}{n} \sum_i \frac{1}{1 + \sum_{\substack{s_i \in S_i \\ s_i \neq x_i}} e^{-\beta[\Phi(\mathbf{x}_{-i}, s_i) - \Phi(\mathbf{x})]}} \\ &= \frac{1}{n} \sum_i \frac{\sum_{\substack{s_i \in S_i \\ s_i \neq x_i}} e^{-\beta[\Phi(\mathbf{x}_{-i}, s_i) - \Phi(\mathbf{x})]}}{1 + \sum_{\substack{s_i \in S_i \\ s_i \neq x_i}} e^{-\beta[\Phi(\mathbf{x}_{-i}, s_i) - \Phi(\mathbf{x})]}}. \end{aligned} \tag{12}$$

Consider now a sequence of profiles $\mathbf{x}$, one for each number n of players. In order to decide if $B(\mathbf{x})$ is a polynomial in n, one needs to look at how the potential values change with n. Towards this aim, it will be useful to describe these potential values as monotone non-decreasing continuous functions $f_1, \ldots, f_k$ of n. Here we say that f_j *describes* $\Phi(\mathbf{x})$ if there are n_0, c_1, c_2 such that for every $n \geq n_0$ we have $\Phi(\mathbf{x}) \in [c_1 f_j(n), c_2 f_j(n)]$. If the number of these functions f_j is unbounded, as in Lemma B.1, then one may be unable to describe $B(\mathbf{x})$ as the inverse of a polynomial or the inverse of a super-polynomial.

This line of reasoning can then bring one to suppose that, in order to prove asymptotic results about metastability is sufficient that potential values can be described by a finite number of functions. Roughly speaking, this means that potential values are concisely representable. Unfortunately, next lemma shows that, even if game is constrained in this way, we can still design a game for which it does not make sense to prove asymptotic results about metastability.

Lemma B.2. *There is a sequence of n-player (potential) games $\mathbf{G}$ that is not asymptotically metastable for every β sufficiently high and every $\varepsilon < \frac{1}{4}$, even if potential values can be described by a finite number of functions in n.*

Proof. As above we will show a sequence of n-player games $\mathbf{G}$ such that for every $0 < \varepsilon < 1/4$, for infinitely many value of n and for each polynomial p in n and each super-polynomial q in n, there is a profile $\mathbf{x}$ of $\mathcal{G}_n$ such that the logit dynamics for $\mathcal{G}_n$ does not converge in time at most $p(n)$ from $\mathbf{x}$ to any $(\varepsilon, q(n))$-metastable distribution, even if the mixing time of the dynamics is larger than $p(n)$.

Let (p_j, q_j) be defined as in the proof of Lemma B.1. Moreover, let f_1 and f_2 two monotone non-decreasing continuous functions in n with $f_1(n) \geq \frac{q_j(n)}{\varepsilon}$ for every j, (e.g. f_1 is exponential in n). Let us denote as n_j a value such that there is $t_{n_j}^{(j)} \in (p_j(n_j), q_j(n_j))$ such that $\frac{\varepsilon}{t_{n_j}^{(j)}} \cdot \frac{n_j f_1(n_j) f_2(n_j)}{f_1(n_j) - f_2(n_j)} - \frac{n_j f_2(n_j)}{f_1(n_j) - f_2(n_j)}$ is an integer. It is not hard to see that for every $n > n_j$, there will be a value $t_n^{(j)}$ in the interval $(p_j(n), q_j(n))$ such that $\frac{\varepsilon}{t_n^{(j)}} \cdot \frac{n f_1(n) f_2(n)}{f_1(n) - f_2(n)} - \frac{n f_2(n)}{f_1(n) - f_2(n)}$ is an integer. Finally, let us define the function $\mathcal{T}$ as follows: $\mathcal{T}(n) = t_n^{(j)}$ for j such that $n_j \leq n < n_{j+1}$. Note that for every p_j and for every $n \geq n_j$, we have $\mathcal{T}(n) = t_n^{(k)} > p_k(n) \geq p_j(n)$, where $k \geq j$ is such that $n_k \geq n < n_{k+1}$. Similarly, for every q_j and for every $n \geq n_j$ we have $\mathcal{T}(n) = t_n^{(k)} < q_k(n) \leq q_j(n)$.

Let now $\mathbf{G}$ be a sequence of n-player potential games such that, for each $\mathcal{G}_n$, each player has exactly two strategies, say 0 and 1. Consider the potential function Φ of $\mathcal{G}_n$ such that for every $t = 0, \ldots, n-2$ and every profile $\mathbf{x}$ wherein exactly t players play strategy 1 we have $\Phi(\mathbf{x}) = n - t$, $\Phi(1, \ldots, 1) = 0$, whereas

$$k_n = \frac{\varepsilon}{\mathcal{T}(n)} \cdot \frac{n f_1(n) f_2(n)}{f_1(n) - f_2(n)} - \frac{n f_2(n)}{f_1(n) - f_2(n)} \tag{13}$$

profiles with a single player playing strategy 1 have potential $-\frac{\log f_2(n)}{\beta}$, and the remaining $n - k_n$ profiles have potential $-\frac{\log f_1(n)}{\beta}$. Observe that only $n + 1$ functions have been used for describing the potential values of the game.

We will show that $B(1, \ldots, 1) = \frac{\varepsilon}{\mathcal{T}(n)}$, and hence the claim follows adopting exactly the same arguments as in the proof of Lemma B.1. Indeed, from (12) it follows that

$$B(1, \ldots, 1) = \frac{1}{n} \left(\frac{k_n}{f_2(n)} + \frac{n - k_n}{f_1(n)} \right),$$

from which the claim follows by simple algebraic manipulations and from (13). □

B.3 Asymptotically Well-Described Games

Lemma B.2 shows that we need an even more restrictive definition of asymptotically well-behaved games. To this aim, it will be useful to look more closely at (12) and at the game defined in the proof of Lemma B.2.

The condition that potential values are described only by a bounded number of functions assures that each term $e^{-\beta[\Phi(\mathbf{x}_{-i}, z_i) - \Phi(\mathbf{x})]}$ can be in turn described by a bounded number of functions. Still, this does not imply that even $B(\mathbf{x})$ can be described by one among a finite number of functions. Indeed, Lemma B.2 designs a game in which these finitely many functions are arranged in infinitely many ways in order to make the bottleneck ratio of the profile $(1, \ldots, 1)$ to change infinitely often so that it is not possible to classify it as a polynomial or as a super-polynomial.

Hence, the natural approach would be to establish that the games for which it makes sense to prove asymptotic results about metastability not only have potential values described by a bounded number of functions, but also for each profile the fraction of neighbors with a given potential value must be described by a bounded number of functions.

Unfortunately, this definition is still insufficient to have meaningful asymptotic results. Indeed, when the number of players increases, new profiles and new subsets of profiles are created. In order to describe these new subsets of profiles, we need to list the potential values and the neighborhood of these subset.

In particular, one may describe an instance similar to the one of Lemma B.1, except that the role that is played by profile $(1,\dots,1)$, will be now played by a "newly created" profile for each n.

Hence, we must avoid that infinitely many "different" profiles are created as n increases. To this aim, we will require that each new profile $\mathbf{x}$ must be *associated* to a profile $\mathbf{x}'$ with a lower number of players, where, roughly speaking, associated means that $\mathbf{x}$ and $\mathbf{x}'$ have the same behavior when compared with the other profiles.

But we still have some issues. Indeed, whereas in the examples discussed above we considered a single profile that is not metastable enough and still hard to leave, we can now consider a larger subset of profiles for which this statement holds. And as discussed for single profiles, we need also to care about every "newly created" subsets of profiles. For this reason, we need to extend the concept of association from profiles to set of profiles, as follows.

Definition B.1. Let $\mathcal{G}$ be a potential game and $\beta > 0$. A subset L of profiles of $\mathcal{G}_n$ is said to be *associated* to a subset L' of profiles of $\mathcal{G}_{n'}$ for a given β if there are monotone continuous functions[11] $f_1,\dots,f_k\colon \mathbb{R}\to\mathbb{R}$, with finite k, such that $\chi(L,p)=f_q(n)$ if and only if $\chi'(L',p)=f_q(n')$, for every $p,q\in\{1,\dots,k\}$, where

- $\chi(L,p) := \left|\left\{\mathbf{y}\in\partial L\colon \frac{\sum_{j=1}^k \upsilon(\mathbf{y},j)e^{-\beta f_j(n)}}{\sum_{j=1}^k \iota(\mathbf{x},j)e^{-\beta f_j(n)}} = f_p(n)\right\}\right|$;
- $\chi'(L',p) := \left|\left\{\mathbf{y}\in\partial L'\colon \frac{\sum_{j=1}^k \upsilon(\mathbf{y},j)e^{-\beta f_j(n')}}{\sum_{j=1}^k \iota(\mathbf{x},j)e^{-\beta f_j(n')}} = f_p(n')\right\}\right|$;
- $\iota(\mathbf{y},j) := |\{\mathbf{z}\in L\colon \Phi(\mathbf{z})-\Phi(\mathbf{y})=f_j(n)\}|$;
- $\iota'(\mathbf{y},j) := |\{\mathbf{z}\in L'\colon \Phi(\mathbf{z})-\Phi(\mathbf{y})=f_j(n')\}|$;
- $\upsilon(\mathbf{y},j) := \left|\left\{\ell\in[n]\colon \frac{\sum_{w=1}^k \eta(\mathbf{y},\ell,w)e^{-\beta f_w(n)}}{\sum_{w=1}^k \gamma(\mathbf{y},\ell,w)e^{-\beta f_w(n)}} = f_j(n)\right\}\right|$;
- $\upsilon'(\mathbf{y},j) := \left|\left\{\ell\in[n']\colon \frac{\sum_{w=1}^k \eta(\mathbf{y},\ell,w)e^{-\beta f_w(n')}}{\sum_{w=1}^k \gamma(\mathbf{y},\ell,w)e^{-\beta f_w(n')}} = f_j(n')\right\}\right|$;
- $\gamma(\mathbf{y},\ell,w) := |\{s\in S_\ell\colon \Phi(s,\mathbf{y}_{-\ell})-\Phi(\mathbf{y})=f_w(n)\}|$;
- $\gamma'(\mathbf{y},\ell,w) := |\{s\in S'_\ell\colon \Phi(s,\mathbf{y}_{-\ell})-\Phi(\mathbf{y})=f_w(n')\}|$;
- $\eta(\mathbf{y},\ell,w) := |\{s\in S_\ell\colon (s,\mathbf{y}_{-\ell})\notin L \text{ and } \Phi(s,\mathbf{y}_{-\ell})-\Phi(\mathbf{y})=f_w(n)\}|$;
- $\eta'(\mathbf{y},\ell,w) := |\{s\in S'_\ell\colon (s,\mathbf{y}_{-\ell})\notin L' \text{ and } \Phi(s,\mathbf{y}_{-\ell})-\Phi(\mathbf{y})=f_w(n')\}|$.

In words, γ and γ' measure the number of profiles differing from $\mathbf{y}$ only in the strategy of player ℓ and whose difference in potential with $\mathbf{y}$ is given by a certain function evaluated in n and n', respectively; η and η' measure the same quantity but restricted only to neighbors of $\mathbf{y}$ outside the set L and L', respectively. Similarly, ι and ι' measure the number of profiles in L whose potential difference with $\mathbf{y}$ is given by a function of n and n', respectively. Roughly speaking, these quantities are measuring the relation between a profile $\mathbf{y}$ and the profiles that are either in L or in the neighborhood of L. Thus, if these quantities for profile $\mathbf{y}\in L$ assume a given value as the corresponding quantities for profile $\mathbf{y}'\in L'$, then this means that $\mathbf{y}$ and $\mathbf{y}'$ have the same relation with other nodes in L and L', respectively, and with their neighborhoods. Since υ (υ', respectively) aggregates γ and η (γ' and η', respectively), and χ (χ', respectively) aggregates ι and υ (ι' and υ', respectively), we have that if χ assumes the same value

[11] Although these functions depend on β, for the sake of readability, we suppress this dependence from the notation. A similar consideration applies to Lemma B.3.

as χ', then profiles in ∂L and $\partial L'$ have the same relation with other profiles in L and L', respectively, and with their neighborhoods.

We observe that in the definition above we do not really need the ratios/difference to be equal to one of the continuous functions. It is indeed sufficient that those quantities are very close to the value of the function of interest. Specifically, we will use the following definition.

Definition B.2. Let $\mathcal{G}$ be a potential game and $\lambda \geq 0$. A subset L of profiles of $\mathcal{G}_n$ is said to be λ-*associated* to a subset L' of profiles of $\mathcal{G}_{n'}$ for a given β if there are monotone continuous functions $f_1, \dots, f_k \colon \mathbb{R} \to \mathbb{R}$, with finite k, such that the definitions of Definition B.1 hold, except that whenever we say that a ratio/difference r is equal to $f_j(x)$, $j \in [k]$ and $x \in \{n, n'\}$ we now only require that $r \in [(1-\lambda)f_j(x), (1+\lambda)f_j(x)]$.

The next fact directly follows from this definition.

Lemma B.3. *Let $\mathcal{G}$ be a potential game and $\lambda \geq 0$. If a subset L of profiles of $\mathcal{G}_n$ is λ-associated to a subset L' of profiles of $\mathcal{G}_{n'}$, with $n' < n$, for a given β then there is a monotone non-deceasing continuous function $F_{L'} \colon \mathbb{R} \to \mathbb{R}$ such that the bottleneck ratio for the logit dynamics for L' is $B(L') = F_{L'}(n')$ and the bottleneck ratio for the logit dynamics for L satisfies $B(L) \in [F_{L'}(n)(1-\lambda'), F_{L'}(n)(1+\lambda')]$, with $\lambda' > 0$.*

Proof. For sake of readability, we prove the claim only for $\lambda = 0$. The extension to different values of λ is immediate. Observe that

$$B(L) = \frac{1}{n} \sum_{\mathbf{x} \in \partial L} \frac{\pi(\mathbf{x}) \sum_{\mathbf{y} \notin L} P(\mathbf{x}, \mathbf{y})}{\pi(L)} = \frac{1}{n} \sum_{\mathbf{x} \in \partial L} \frac{\sum_i \frac{\sum_{s \in S_i \colon (s, \mathbf{x}_{-i}) \notin L} e^{-\beta \Phi(s, \mathbf{x}_{-i})}}{\sum_{z \in S_i} e^{-\beta \Phi(z, \mathbf{x}_{-i})}}}{\sum_{\mathbf{y} \in L} e^{-\beta [\Phi(\mathbf{y}) - \Phi(\mathbf{x})]}}$$
$$= \frac{1}{n} \sum_{\mathbf{x} \in \partial L} \frac{\sum_i \frac{\sum_{s \in S_i \colon (s, \mathbf{x}_{-i}) \notin L} e^{-\beta [\Phi(s, \mathbf{x}_{-i}) - \Phi(\mathbf{x})]}}{\sum_{z \in S_i} e^{-\beta [\Phi(z, \mathbf{x}_{-i}) - \Phi(\mathbf{x})]}}}{\sum_{\mathbf{y} \in L} e^{-\beta [\Phi(\mathbf{y}) - \Phi(\mathbf{x})]}}.$$

Since L is associated to L' for the given β, then there are $f_1, \dots, f_k$ as in Definition B.1. Moreover,

$$\sum_{s \in S_i \colon (s, \mathbf{x}_{-i}) \notin L} e^{-\beta [\Phi(s, \mathbf{x}_{-i}) - \Phi(\mathbf{x})]} = \sum_{j=1}^{k} \eta(\mathbf{x}, i, j) e^{-\beta f_j(n)}$$
$$\sum_{z \in S_i} e^{-\beta [\Phi(z, \mathbf{x}_{-i}) - \Phi(\mathbf{x})]} = \sum_{j=1}^{k} \gamma(\mathbf{x}, i, j) e^{-\beta f_j(n)}$$
$$\sum_{\mathbf{y} \in L} e^{-\beta [\Phi(\mathbf{y}) - \Phi(\mathbf{x})]} = \sum_{j=1}^{k} \iota(\mathbf{x}, j) e^{-\beta f_j(n)}.$$

Hence, we have that

$$B(L) = \frac{1}{n} \sum_{\mathbf{x} \in \partial L} \frac{\sum_i \frac{\sum_{j=1}^{k} \eta(\mathbf{x}, i, j) e^{-\beta f_j(n)}}{\sum_{j=1}^{k} \gamma(\mathbf{x}, i, j) e^{-\beta f_j(n)}}}{\sum_{j=1}^{k} \iota(\mathbf{x}, j) e^{-\beta f_j(n)}} = \frac{1}{n} \sum_{\mathbf{x} \in \partial L} \frac{\sum_{j=1}^{k} \upsilon(\mathbf{x}, j) e^{-\beta f_j(n)}}{\sum_{j=1}^{k} \iota(\mathbf{x}, j) e^{-\beta f_j(n)}} = \frac{1}{n} \sum_{j=1}^{k} \chi(L, j) e^{-\beta f_j(n)}.$$

Similarly, we achieve that

$$F_{L'}(n') = B(L') = \frac{1}{n'} \sum_{j=1}^{k} \chi'(L', j) e^{-\beta f_j(n')},$$

from which the claim immediately follows. □

We are now ready to redefine the concept of asymptotically well-behaved games. In order to distinguish this definition from Definition 4.2, we call these game *asymptotically well-described*.

Definition B.3. A potential game $\mathcal{G}$ is *asymptotically well-described* for β if there is n_0 such that for every $n \geq n_0$, every subset L of profiles of $\mathcal{G}_n$ is λ-associated, $\lambda = O(1)$, with a subset L' of profiles of $\mathcal{G}_{n_0}$ for the given β.

We will call the constant n_0 in the above definition the *asymptotic basis* of the game.

We next show that if a game is asymptotically well-described, then it is also asymptotically well-behaved.

Lemma B.4. *Let $\beta > 0$ and let $\mathbf{G}$ be a sequence of n-player potential games asymptotically well-described for β, with n_0 as its asymptotic basis. Then, there is a polynomial p and a super-polynomial q such that for every $n \geq n_0$ and for every $L' \subseteq \mathcal{S}_n$, either $B(L') \geq \frac{1}{p(n)}$ or $B(L') \leq \frac{1}{q(n)}$.*

Proof. According to the definition of asymptotically well-behaved game and to Lemma B.3, there is a subset $L \subseteq S = S_1 \times \cdots \times S_{n_0}$, a simple function F_L, and a small constant λ such that $B(L') \in [F_L(n)(1-\lambda), F_L(n)(1+\lambda)]$. Thus $B(L')$ can be upper-bounded by a polynomial if F_L is a polynomial and it can be lower-bounded by a super-polynomial if F_L is a super-polynomial. Then, it is only left to show that it is always possible to distinguish if F_L is either a polynomial or a super-polynomial. Indeed, since there are a finite number (specifically, at most m^{n_0}) of such functions F_L, we can partition the subsets of S in two (possibly empty) subsets S' and S'' such that for every L in S', $F_L(n) \geq 1/p(n)$ and, for every L in S'', $F_L(n) \leq 1/q(n)$, for every $n \geq n_0$ and for p polynomial and q super-polynomial. □

C Asymptotically Well-Classified Games

The definition of asymptotically well-behaved games turns out to be a bit too stringent for our techniques. Indeed, there are games for which the definition does not hold, but it is still possible to prove asymptotic metastability through our framework.

Consider indeed the following game: the *pure coordination game* is an n-player game where players have the same strategy set A and each player is happy when all players adopt the same strategy and unhappy otherwise. It is not hard to see that this is a potential game with potential function Φ defined as $\Phi(\mathbf{x}) = 1$ in every profile $\mathbf{x}$ where players do not coordinate and $\Phi(\mathbf{x}) = 0$ for the remaining profiles $\mathbf{x}$.

Specifically, in [5], it has been considered the case in which each agent can choose between two strategies, namely $+1$ and -1; each agent has utility 1 if all the players adopt the same strategy and utility 0 otherwise. The mixing time of the logit dynamics for these games is polynomial for $\beta = \mathcal{O}(\log n)$ and super-polynomial otherwise. Auletta et al. [5] show asymptotic metastability for every $\beta = \omega(\log n)$. Next lemma proves also that n-player pure coordination games are partitioned by the logit dynamics.

Lemma C.1. *Let $\mathbf{G}$ be a sequence of n-player pure coordination game. Then $\mathbf{G}$ is partitioned by the logit dynamics for every $\beta = \omega(\log n)$.*

Proof. Fix n and consider the following subsets of the set S of profiles of $\mathcal{G}_n$: $R_1 = \{\mathbf{p}\}, R_2 = \{\mathbf{m}\}$ and $R_3 = \{+1, -1\}^n \setminus \{\mathbf{p}, \mathbf{m}\}$, where $\mathbf{p} = (+1)^n$ and $\mathbf{m} = (-1)^n$. As showed in [5], the bottleneck ratio of these subsets is super-polynomial for every $\beta = \omega(\log n)$. Moreover, the mixing time of the chains restricted to R_1 and R_2 is trivially polynomial. As for R_3, observe that the stationary distribution of the restricted chain is very close to the stationary distribution of a lazy random walk on an n-dimensional hypercube, whose mixing time is known to be polynomial (see, e.g., [25]).

Also let us consider the following partition of S: $T_i = R_i$ for $i = 1, 2, 3$ and N is empty. Clearly, T_1 and T_2 satisfy the property required by the definition of partitioned chains. This holds also for

T_3. Indeed, consider a birth and death chain (see, e.g., Section 2.5 in [25]) defined on the state space $\{0, 1, \ldots, m\}$ with transition probability:

$$p_0 = q_m = r_0 = r_m = \frac{1}{2}, q_0 = p_m = 0; \qquad p_i = \frac{m-i}{4m}, q_i = \frac{m+i}{4m}, r_i = \frac{1}{2}, \text{ for } i = 1, \ldots, m-1;$$

where p_i is the probability of going from state i to state $i+1$, q_i is the probability of going from state i to state $i-1$ and r_i is the probability to stay in state i.

If $m = n/2$[12], then the above birth an death chain can be seen as the projection of our chain, where the state $\mathbf{x}$ of our chain is projected to the state i of the birth and death chain, such that the minimum among the zeros and the ones in $\mathbf{x}$ is $\frac{n}{2} - i$. Hence, the expected hitting time of either $\mathbf{p}$ or $\mathbf{m}$ is equivalent to the expected hitting time of m. It is then easy to check that the expected hitting time of this state is super-polynomial in n from every starting state (see, e.g., Section 2.5 in [25]). The claim finally follows by a simple application of Markov's inequality. □

Still, the following lemma proves that it is not asymptotically well-behaved.

Lemma C.2. *The pure coordination game is not asymptotically well-behaved.*

Proof. Consider p_j, q_j as in Lemma B.1. Let n_j be the first value of n such that there is an integer $t_n^{(j)} \in \{2, \ldots, n-2\}$ for which

$$T = 2\varepsilon \cdot \frac{\sum_{i=2}^{t_n^{(j)}} \binom{n}{i}}{\binom{n}{2} + \binom{n}{t_n^{(j)}}} \in [p_j(n), q_j(n)] .$$

Note that n_j always exists since the size of the interval $[p_j(n), q_j(n)]$ goes to infinity with n. Moreover, if such an integer $t_n^{(j)}$ exists for $n = n_j$, then it exists also for every $n > n_j$.

Now, for every n, we consider the set $L_n^{(k)}$ that contains all profiles with at least two and at most $t_n^{(k)}$ players adopting strategy -1, where k is such that $n_k \leq n \leq n_{k+1}$. Observe that by definition of $t_n^{(j)}$ all profiles in $L_n^{(k)}$ and their neighbors have the same potential. Hence, the bottleneck ratio of $L_n^{(k)}$ is exactly $\frac{\varepsilon}{T}$. Thus, for every j there always exists a subset such that its bottleneck is larger than p_j and smaller than q_j. That is, we cannot find a polynomial p and a super-polynomial q that bound the bottleneck ratio of all subsets of profiles of the game. Then, we can conclude that this game is not asymptotically well-behaved. □

This example shows a weakness of our definition of asymptotically well-behaved games. Specifically, in Definition 4.2 we requested that for these games it holds that we can asymptotically classify the bottleneck ratio of each subset of profiles as either polynomial or super-polynomial. More specifically, we can assume there are two functions p at most polynomial in the input and q at least super-polynomial in the input such that for each β and each subset A the bottleneck ratio $B(A)$ can be bounded by functions that depends on either p or q. An equivalent viewpoint would be to see a sequence of n-player potential games as a class to which a kind of oracle is attached that distinguishes between polynomial and super-polynomial bottleneck ratios for every fixed β. Formally, given a sequence of n-player potential games $\mathbf{G}$ and fixed $\beta \geq \beta_0$, this oracle can be described as follows: when it is queried about the bottleneck ratio of a subset A with n players its answer states that the bottleneck ratio is either i) at most polynomial if it is lower-bounded by $1/p(n)$; or ii) at least super-polynomial if it is upper-bounded by $1/q(n)$.

In turn, we proved that this oracle is sufficient to prove that the game can be partitioned by the logit dynamics, and thus enjoys asymptotic metastability. However, the example above states that this oracle

[12] Here, we are assuming n is even. The case for odd n is similar.

is not necessary. A more careful look to the pure coordination game discussed above actually highlights that to prove asymptotic metastability of a game with respect to a polynomial p and a super-polynomial q we do not need the behavior of each subset to be classified. That is, we can allow some subsets of profiles to have bottleneck ratio in between the inverse of q and the inverse of p. In this case, we will say that the subset of profiles is *unclassified*.

In this way, we can weaken our definition, by presenting a condition that describes which class of subsets is sufficient to classify in order to have that the sets returned by Algorithm 4.1 enjoy the properties required by the definition of partitioned chains. In particular, we define the class of *asymptotically well-classified games* as follows.

Definition C.1 (Asymptotically well-classified games)**.** A sequence of n-player potential games $\mathbf{G}$ is asymptotically well-classified for $\beta \geq \beta_0$ if for $\varepsilon > 0$ there exist a pair of functions p at most polynomial and q at least super-polynomial, that for each n, satisfy the following conditions:

1. $q(n) \leq \max_{L\colon \pi(L)\leq 1/2} B^{-1}(L)$;
2. for each R_i computed by $\mathcal{A}_{p,q}$ and for every $L \subset R_i$ such that $\pi_{R_i}(L) \leq 1/2$, if $B_{R_i}(L) < 1/p(n)$, then both $B(L)$ and $B(R_i \setminus L)$ are not unclassified;
3. for each subset $L \subseteq N$, N being as at the end of the algorithm $\mathcal{A}_{p,q}$, such that $\pi(L) \leq 1/2$, $B(L)$ is not unclassified.

By careful looking at their proofs, one can check that Proposition 4.1 and Proposition 4.2 continue to hold even if we substitute asymptotically well-behaved potential games with asymptotically well-classified ones.

D Spectral properties of the logit dynamics

We next give other interesting spectral results about the transition matrix generated by the logit dynamics. In particular, by using a matrix decomposition similar to the one adopted in the proof of Lemma 4.2 we can prove the following propositions. (We remark that results in this section do not need to assume that the chain is reversible and indeed apply to every strategic game and not only to potential games.)

Proposition D.1. *Let $\mathcal{G}$ be a game with profile space S and let P be the transition matrix of the logit dynamics for $\mathcal{G}$. The trace of P is independent of β.*

Proof. For every i and for every $\mathbf{z}_{-i}$ consider the transition matrices $P_{i,\mathbf{z}_{-i}}$ defined in (6), with $L = S$. Let $S_{i,\mathbf{z}_{-i}} = \{(\mathbf{z}_{-i}, s_i) \mid s_i \in S_i\}$. Observe that for every $\mathbf{x} \in S_{i,\mathbf{z}_{-i}}$ we have $P_{i,\mathbf{z}_{-i}}(\mathbf{x},\mathbf{x}) = 1 - \sum_{\mathbf{y}\in S_{i,\mathbf{z}_{-i}},\mathbf{y}\neq\mathbf{x}} P(\mathbf{x},\mathbf{y})$. Hence, the trace of $P_{i,\mathbf{z}_{-i}}$ is

$$\sum_{\mathbf{x}\in S_{i,\mathbf{z}_{-i}}} P_{i,\mathbf{z}_{-i}}(\mathbf{x},\mathbf{x}) = |S_i| - \sum_{\mathbf{x}\in S_{i,\mathbf{z}_{-i}}} \sum_{\mathbf{y}\in S_{i,\mathbf{z}_{-i}},\mathbf{y}\neq\mathbf{x}} P(\mathbf{x},\mathbf{y}).$$

Since all non-zero elements in a column of $P_{i,\mathbf{z}_{-i}}$ are the same we also have

$$P_{i,\mathbf{z}_{-i}}(\mathbf{x},\mathbf{x}) = \frac{1}{|S_i|-1} \sum_{\mathbf{y}\in S_{i,\mathbf{z}_{-i}},\mathbf{y}\neq\mathbf{x}} P(\mathbf{y},\mathbf{x}).$$

By setting $C = \sum_{\mathbf{x}\in S_{i,\mathbf{z}_{-i}}} \sum_{\mathbf{y}\in S_{i,\mathbf{z}_{-i}},\mathbf{y}\neq\mathbf{x}} P(\mathbf{x},\mathbf{y}) = \sum_{\mathbf{x}\in S_{i,\mathbf{z}_{-i}}} \sum_{\mathbf{y}\in S_{i,\mathbf{z}_{-i}},\mathbf{y}\neq\mathbf{x}} P(\mathbf{y},\mathbf{x})$, we have

$$|S_i| - C = \frac{C}{|S_i|-1} \Longrightarrow C = |S_i| - 1,$$

and thus, the trace of $P_{i,\mathbf{z}_{-i}}$ is always 1, regardless of β. The proposition follows since the trace of P is exactly the sum of the traces of all $P_{i,\mathbf{z}_{-i}}$'s. □

The proposition above says that if there exists an eigenvalue of P that gets closer to 1 as β increases, then there are other eigenvalues that get smaller: this is very promising in the tentative to characterize the entire spectrum of eigenvalues of P, necessary to use powerful tools such as the well-known *random target lemma* (see, e.g., [25]).

In order to prove our last characterization of the transition matrix generated by the logit dynamics, we prove the following lemma which gives a lower bound on the probability that the strategy profile is not changed in one step of the logit dynamics for a generic game.

Lemma D.1. *Let $\mathcal{G}$ be a game with profile space S and let P be the transition matrix of the logit dynamics for $\mathcal{G}$. Then for every $\mathbf{x} \in S$ we have that*

$$P(\mathbf{x},\mathbf{x}) = \sum_i P\Big((\mathbf{x}_{-i}, s_i^\star), \mathbf{x}\Big),$$

where $s_i^\star \neq x_i$ is an arbitrary strategy of player i.

Proof. Observe that

$$\begin{aligned} P(\mathbf{x},\mathbf{x}) = 1 - \sum_{\mathbf{y}\in N(\mathbf{x})} P(\mathbf{x},\mathbf{y}) &= \sum_i \left(\frac{1}{n} - \sum_{\mathbf{y}\in N_i(\mathbf{x})} P(\mathbf{x},\mathbf{y})\right) \\ &= \sum_i \frac{1}{n}\left(1 - \sum_{\mathbf{y}\in N_i(\mathbf{x})} \frac{e^{\beta u_i(\mathbf{y})}}{e^{\beta u_i(\mathbf{x})} + \sum_{\mathbf{z}\in N_i(\mathbf{x})} e^{\beta u_i(\mathbf{z})}}\right) = \sum_i \frac{1}{n} \frac{e^{\beta u_i(\mathbf{x})}}{e^{\beta u_i(\mathbf{x})} + \sum_{\mathbf{z}\in N_i(\mathbf{x})} e^{\beta u_i(\mathbf{z})}}. \end{aligned}$$

The proof concludes by observing that for every i and for every $s_i^\star \in S_i$, we have

$$P\Big((\mathbf{x}_{-i}, s_i^\star), \mathbf{x}\Big) = \frac{1}{n} \frac{e^{\beta u_i(\mathbf{x})}}{e^{\beta u_i(\mathbf{x})} + \sum_{\mathbf{z}\in N_i(\mathbf{x})} e^{\beta u_i(\mathbf{z})}}. \qquad \square$$

Lemma D.1 allows us to calculate the determinant of P.

Proposition D.2. *Let $\mathcal{G}$ be a game with profile space S and let P be the transition matrix of the logit dynamics for $\mathcal{G}$. The determinant of P is 0.*

Proof. It is well-known that a matrix in which one row can be expressed as a linear combination of other rows has determinant zero. In this proof, we fix a profile $\mathbf{x}$ and show that the row of P corresponding to $\mathbf{x}$ can be obtained as a linear combination of other rows of the matrix. For each player i, fix a strategy $s_i^\star \in S_i$ such that $s_i^\star \neq x_i$. Let us denote with S^j, $j = 0, \ldots, n$, the set of profiles $\mathbf{y} \in S$ obtained from $\mathbf{x}$ by selecting j players $i_1, \ldots, i_j$ and setting their strategies to $s_{i_1}^\star, \ldots, s_{i_j}^\star$, respectively. Notice that $\mathbf{x}$ belongs to S^0. By construction, for every profile $\mathbf{z} \in S^j$, $z_i \in \{x_i, s_i^\star\}$. Now, for $i = 1, \ldots, n$, consider the profile obtained from $\mathbf{z}$ by changing $z_i = x_i$ into $s_i^\star$ or viceversa. Note that there are n of such profiles which are neighbors of $\mathbf{z}$ and all contained in the sets S^{j-1} and S^{j+1}. We claim that for every $\mathbf{y} \in S$

$$P(\mathbf{x},\mathbf{y}) = \sum_{j=1}^{n} (-1)^{j+1} \sum_{\mathbf{z}\in S^j} P(\mathbf{z},\mathbf{y}). \tag{14}$$

In order to prove the claim we distinguish three cases:

1. Let $H(\mathbf{x},\mathbf{y}) > 1$ (and thus $P(\mathbf{x},\mathbf{y}) = 0$): if there exists $j \in \{0, \ldots, n\}$ such that $\mathbf{y} \in S^j$, then the r.h.s. of (14) becomes $\pm\Big(P(\mathbf{y},\mathbf{y}) - \sum_i P\big((\mathbf{y}_{-i}, s_i^\star), \mathbf{y}\big)\Big) = 0$, from Lemma D.1; if $\mathbf{y} \notin \bigcup_{j=0}^{n} S^j$, then consider a profile $\mathbf{z} \in S^j$, for some $j = 1, \ldots, n$, such that $\mathbf{z}$ differs from $\mathbf{y}$

only in the strategy of player k: if no such profile exists, then the r.h.s. of (14) is 0; otherwise, let us assume w.l.o.g. $z_k = x_k$ (the case $z_k = s_k^\star$ can be handled similarly), then the profile $\mathbf{z}' = (\mathbf{z}_{-k}, s_k^\star)$ is a neighbor of $\mathbf{y}$, belongs to the set S^{j+1} and $P(\mathbf{z}, \mathbf{y}) = P(\mathbf{z}', \mathbf{y})$: hence, this two profiles delete each other in the r.h.s. of (14), giving the aimed result.

2. Let $\mathbf{x}, \mathbf{y}$ differ in the strategy adopted by the player k: if $\mathbf{y} \in S^1$, then the r.h.s. of (14) becomes $P(\mathbf{y}, \mathbf{y}) - \sum_{i \neq k} P\big((\mathbf{y}_{-i}, s_i^\star), \mathbf{y}\big) = P(\mathbf{x}, \mathbf{y})$, from Lemma D.1; if $\mathbf{y} \notin \bigcup_{j=0}^{n} S^j$, then, as above, all profiles in $\bigcup_{j=0}^{n} S^j$ that differ from $\mathbf{y}$ only in one player $i \neq k$ delete each other in the r.h.s. of (14): thus, the only element that survives in the r.h.s. of (14) is $P\big((\mathbf{y}_{-k}, x_k), \mathbf{y}\big) = P(\mathbf{x}, \mathbf{y})$.

3. If $\mathbf{x} = \mathbf{y}$, then the r.h.s. of (14) becomes $\sum_{i \neq k} P\big((\mathbf{y}_{-i}, s_i^\star), \mathbf{y}\big) = P(\mathbf{x}, \mathbf{x})$, from Lemma D.1. □

Since, as observed above, the logit dynamics for potential games defines a reversible Markov chain, Lemma 4.2 and Proposition D.2 imply that the last eigenvalue of the logit dynamics for these games is exactly 0. (Note that in [3] is only stated the last eigenvalue is non-negative.) Moreover, from the proof above, it turns out that an eigenvector of such zero eigenvalue is given by the function $f \colon S \to \mathbb{R}$ defined as

$$f(\mathbf{w}) = \begin{cases} -1, & \text{if } \mathbf{w} \in S^j \text{ and } j \text{ is even;} \\ 1, & \text{if } \mathbf{w} \in S^j \text{ and } j \text{ is odd;} \\ 0, & \text{otherwise;} \end{cases}$$

where the sets S^j's are defined as in the above proof for some fixed profile $\mathbf{x}$.